AD NUMBER

AD869857

NEW LIMITATION CHANGE

TO

Approved for public release, distribution unlimited

FROM

Distribution authorized to U.S. Gov't. agencies and their contractors; Administrative/Operational Use; APR 1970. Other requests shall be referred to Commanding Officer, Edgewood Arsenal, Attn: SMUEA-TSTI-T, Edgewood Arsenal, MD 21010.

AUTHORITY

USAEA notice, 8 Jul 1970

CIDS No. 7

QUERY FORMULATION AND ENCODING

Status Report

by

Margaret Milne

Paul R. Weinberg

30 April 1970

Distribution Statement

This document is subject to special export controls and each
transmittal to a foreign government or foreign national may
be made only by prior approval of the Commanding Officer,
Edgewood Arsenal, ATTN: SMUEA-TSTI-T, Edgewood Arsenal,
Maryland   21010

Disclaimer

The findings in this report are not to be construed as an
official Department of the Army position unless so designated
by other authorized documents.

Disposition

Destroy this report when no longer needed.   Do not return it
to the originator.

CIDS No. 7

QUERY FORMULATION AND ENCODING

Status Report

by

Margaret Milne
Paul R. Weinberg
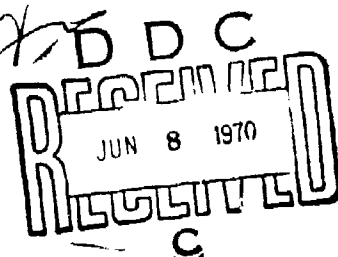
30 April 1970

DEPARTMENT OF THE ARMY
EDGEWOOD ARSENAL

Technical Data & Value Engineering Management Office
Technical Support Directorate
Edgewood Arsenal, Maryland 21010

Contract DAAA15-69-C-0140
Task 2P062101A72702

UNIVERSITY OF PENNSYLVANIA

Philadelphia, Pennsylvania 19104

FOREWORD

The work described in this report was authorized under task 2P062101A72702, Army Chemical Information and Data Systems (U). The work was started in July 1964 and is continuing. The information contained in this report represents part of the work accomplished primarily during the calendar year 1969.

The information in this document has not been cleared for release to the general public.

This publication describes and illustrates the rules for encoding queries addressed to the initial model of an operational chemical information and dat system (CIDS). The teletype command language for submitting these queries to the system was described in an earlier publication by Sherr (1) and is included in this document as Appendix A. The coding rules constitute an update of the method described by Weinberg (2), which applied to the earlier experimental system. As a result of extensive experimentation with that system, the capabilities of CIDS have been expanded for the model system, thus necessitating the current revision. In addition to describing the encoding and input of retrieval demands, this document discusses certain features of system operation and use to assist in the formulation of maximally efficient queries.

The chemical features which provide the basis for a search of the CIDS file of compounds are collectively referred to as the CIDS search components. The complete collection of components, along with the code by which each is represented, is contained in the CIDS No. 6 report entitled Handbook of CIDS Chemical Search Components (3). The current document discusses the encoding of these components into a valid query and includes the procedure for initiating and modifying file search in response to that query. The CIDS No. 6 report is thus corequisite to the current document for use of the real-time model operational system.

A query coding form has been devised on which the user's question, the pertinent search components, and the fully encoded query are recorded. To simplify reference to the rules during the actual formulation and encoding of a query, the topics in this document have been organized to parallel the coding form. Thus certain sections of this document are of primary interest to the chemist selecting the components appropriate to a given query, while other sections dealing with the encoding of these components are of major concern to computer personnel.

The final section of this document contains the coding forms for some twenty sample questions exactly as tnese forms are completed for the operational system. Accompanying each example is an explanation of the strategy emplcyed in the search. Besides providing complete illustrations of query encoding, these examples are designed to suggest to the user a number of techniques for maximizing search efficiency.

TABLE OF CONTENTS

QUERY FORMULATION AND ENCODING

1. INTRODUCTION

The initial model of an operational Chemical Information and Data System (CIDS) selectively retrieves compounds from the CIDS file on the basis of the presence or absence of specified chemical characteristics. The features which can be specified encompass both structure and molecular formula and have been described, categorized, and illustrated in an earlier CIDS publication (3). The primary purpose of the current document is to prescribe for the encoding of these specifications into the formal query used as input to CIDS.

In order to utilize CIDS most efficiently only the minimum set of features which distinguish the desired family of compounds should be stipulated. Determination of this "minimum set" requires, in addition to sound chemical judgment, some basic understanding of the search system. Section 2 of this document therefore contains a summary of the CIDS retrieval process, with emphasis on the primary search tool, the keys. In Section 3 the details for encoding queries on the standard coding form are provided along with discussions of the use of the various search components. The last section (Sec. 4) consists entirely of fully explained examples, and provides the reader with illustrations of the principles discussed in Sections 2 and 3.

2. QUERY FORMULATION AND THE RETRIEVAL PROCESS*

Query† formulation refers to the selection of a suitable combination of CIDS search components to retrieve the family of compounds demanded in a question addressed to the system. To be maximally efficient, a query must employ that combination of components which most effectively identifies true responses in the shortest amount of time. Sometimes it proves most efficient to use a less restrictive query allowing the output to include a small number of false retrievals, rather than expend a substantial amount of computer time to eliminate these few invalid responses. In other cases, a single question can be most economically answered by using two queries each of which retrieves a portion of the desired family.

Three types of search components are employed in CIDS: the keys, the molecular formula statement, and the atom-by-atom search. Each key is associated with a list of the locations in the CIDS file of every compound‡ that contains the chemical feature tagged by that key. If a key is assigned to a compound more than once, the file location of that compound appears that number of times in the keylist. By combining appropriate keylists with multipliers (1,2,3,...) and the connectors AND, OR and NOT, a list of the locations of compounds possessing required characteristics (the so-called accession list) is obtained. The operation of these multipliers and connectors on two keylists A and B is explained below.

---

* For a more detailed description of the search system, see the publication by Powers (4).

† In CIDS terminology, the question is the user's description of the features that characterize the required family of compounds, whereas the query refers to the formal statement of the combination of CIDS search components that functions in the retrieval.

‡ Note that the keylist contains the locations of total compounds; it does not reference individually the structures that constitute the total compound. This means that if a feature occurs more than once in a compound, all occurrences are not necessarily part of the same structure. (Example: Demanding two secondary alcohol groups retrieves both

```
        O
        ||
C-C-C-C-O-C-C-C      and      | Ph-C-C-C-N |   ·   | C-C-C-O |
    |       |                  |     |     |         |   |    |
    O       O                  |     O     |         |   O    |
```

8
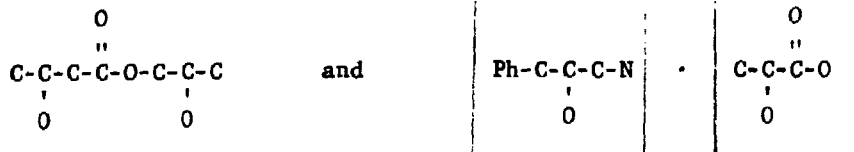
A AND B	retrieves all compounds common to lists A and B, i.e., all compounds containing both feature A and feature B.
A OR B	retrieves all compounds appearing in either list, i.e., all compounds that contain either feature A or feature B or both.
A AND NOT B	retrieves all compounds contained in list A except those that also appear in list B, i.e., all compounds that contain feature A but do not contain feature B.
3A	retrieves all compounds that appear in list A at least three times, i.e., all compounds that contain feature A three or more times.

As illustrated above, demanding "nA" retrieves compounds assigned key A n or more times. If the number of assignments of A to each retrieved compound is required to fall in some specified range, say 3 to 5, it is necessary to demand "3A AND NOT 6A," which in effect identifies compounds having three or more assignments of A and then eliminates those having six or more assignments. This same strategy is used to retrieve compounds assigned key A exactly n times by demanding "nA AND NOT n+1A."

The retrieval demands are frequently more complex than two keys connected by a single conjunction. For example, the user may be interested in compounds having two or more assignments of either of two keys A or B. In either case, he also requires that C occur, but not D. These demands can be stated,

$$(2A \text{ OR } 2B) \text{ AND } C \text{ AND NOT } D$$

which is evaluated internally as follows:

To eliminate parentheses, the above is expanded to the form

2A AND C AND NOT D

OR

2B AND C AND NOT D

consisting of two expressions "ORed" together, each of which is referred to as a minterm. Each minterm is evaluated separately and then the results of all the minterms are combined. Since negating a keylist eliminates those compound locations from the responses to the non-negated keys, at least one non-negated

9

key must appear in each minterm. The following expressions are invalid for the reasons cited.

NOT A AND NOT B   (minterm contains only negated keys)

A OR NOT B   (first minterm correct, but second contains only negated keys)

NOT (A OR B)   (both minterms contain only negated keys)

The overall efficiency of a search is affected not only by the number and list length of the individual keys employed, but also by the order in which they appear in the input.* Consider a query to retrieve all compounds which have been assigned each of three keys: key A of short list length and keys B and C of longer list length. The retrieval system first forms an intermediate list of the compounds common to two of the keys and then intersects this list with the third key list. If the longer list keys B and C are intersected first, considerable computer time is required and the resulting intermediate list may also be lengthy. However, first intersecting the short key list of A with the shorter of B or C not only requires less time, but also guarantees a short intermediate list (since the intermediate list can be no longer than the shorter of the lists intersected for form it). Since the order of intersection and negation of the key lists depends in part† on the order in which the keys in the query are input, the key having the shortest list length - i.e., the key that tags the least common chemical feature - should appear first in the input, with the remaining keys in expected order of increasing list length.

Satisfaction of the keys is always the first step in query processing. The compounds located in the file in response to the keys may be printed out directly, or else additional restrictions can be imposed by the Molecular Formula Statement and/or the Atom-by-Atom Search. The formula statement is

---

* "... the order in which (the keys are)...input" is the order in which their keydefinitions appear (described in Section 3.5.2.1). It does not refer to their order of appearance in the KEYS logical statement (described in Section 3.5.2.2).

† Negation of keylists in a minterm necessarily occurs after all intersections, even if the negated keys are input first.

a very rapid technique for limiting the counts and types of elements that appear in the Hill and/or addend molecular formulas of all retrieved compounds. It enables the user to define a range of possible formulas, a specification which can not be economically stated through the molecular formula keys.

The A/A search allows the user to search for structural features which are not specifically tagged by the structural fragment keys. More than one such fragment can be specified, using multipliers and the connectors AND, OR and NOT to form minterms essentially* as described above. As the name atom-by-atom search implies, each atom of a potential response must be examined individually to determine if it does in fact correspond to an atom in the fragment demanded in the query. Such searches are therefore quite lengthy and it is frequently more efficient to effect the final discrimination simply by visual examination of the compounds retrieved by the keys and formula statement alone.

In the working system, a chemist familiar with the search components lists the keys to be employed in the search, and specifies the restrictions to be imposed by the Molecular Formula Statement and the structure(s) to be satisfied by A/A search. These demands are then encoded, and the typing of input can proceed. The form on which the initial question and all pertinent specifications and coding are recorded is illustrated and described in the following section. Systematic completion of this form according to the rules contained in subsequent sections of this document results in the fully encoded query ready for input.

---

* While every keys minterm must include a non-negated key, it is permitted to negate all of the fragments referenced in a structure minterm.

11

3. QUERY CODING FORM

The form on which queries addressed to CIDS are encoded is illustrated on pages 13 to 15.[*] In addition to consolidating the chemical and computer data relative to a given query, this single coding form has been designed so that a teletype operator can input directly from these sheets. The form is divided into five parts as follows:

    I   User's Question

    II   Encoded Query

    III Molecular Formula and Structural Fragment Keys

    IV   Molecular Formula Statement

    V   Atom-by-Atom Search

Parts I, III, IV and a portion of V contain the chemical specification of retrieval demands, while the remaining parts contain the encoding of these demands for input. In the discussions of each part which follow, the rules for encoding Part II are delayed until last so that the various parts are explained in the sequence in which they are completed in actual use.

---

[*] In actual practice, the form will be presented on a single sheet and will provide space for recording search statistics.

Query name:

I. QUESTION:

Structural Representation	Molecular Formula Specifications

II. ENCODED QUERY:

Query name:

Keydefinitions: (Section III.)

KEYS =

FORMULA

DEFINE STRUCTURE   (Section V.)

STRUCTURE =

END

III.

KEYS			
User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)			
Acyclic-Cyclic (A-C)			
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)			
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
/////						

14

V. ATOM-BY-ATOM SEARCH

Structure(s):

Structure name	Structurespecification	No. of Occurrences

15

3.1 Part I: User's Statement of the Query

In Part I (illustrated below) a user desiring to query CIDS states his question using whatever conventional chemical terminology and/or symbolism he prefers. No detailed knowledge of the CIDS search components is required to complete this section. All that is necessary is a precise definition of the chemical characteristics that all true responses must possess. No encoding is done in this part, which serves only to supply a convenient reference to the original question when filling out the remaining sections of the coding form, and to consolidate for future reference the chemical and computer data pertinent to a given query.

Illustration: Query Coding Form, Part I

I.	QUESTION:	
	Structural Representation	Molecular Formula Specifications

CIDS recognizes that many questions can be expressed more clearly in terms of a complete or partial structure and/or a complete or partial molecular formula common to all true responses. Therefore the query coding form provides sections for stipulating structural and/or molecular formula specifications to supplement - or even replace - the verbal statement of the question. Structures need not be drawn in any particular format as long as the representation employed is complete and unambiguous. The inclusion of a structural representation and/or a molecular formula specification whenever possible is encouraged because, by circumventing language problems, they constitute the most concise and precise way of stating a question.

It is emphasized that query formulation is not necessarily the direct encoding of the specifications stated in Part I. These specifications simply provide a complete and accurate description of the desired family of compounds. The resulting query or queries will specify only the minimum set of chemical features necessary to retrieve that family, regardless of whether or not these features were explicitly mentioned in the original statement of the question.

3.2 Part III: Specification of the Keys

On the basis of the information provided by the user in Part I, the query or queries* that most efficiently retrieve all true answers are formulated using the keys and, as required, the formula statement and/or an atom-by-atom search. In Part III of the coding form for each query, the keys required to tag each response are specified. The final encoding of these keys for input is completed in Part II of the coding form as described in Section 3.5. For the user's convenience, Part III contains a checklist of the various types of molecular formula and structural fragment keys.

Illustration: Query Coding Form, Part III

III.	KEYS			
	User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
	Mol. Formula (MF)			
	Acyclic-Cyclic (A-C)			
	Extracyclic (EC)			
	Number of Cyclic Nuclei (NCN)			
	Cyclic Nuclei: non-H Attmts. (DACN)			
	Generic Cyclic Nuclei (GCN)			
	Specific Cyclic Nuclei (SCN)			
	Specific Functional Group (FG)			
	Nonspec. Diatomics (ND)			
	Nonspec. Monatomics (NM)			
	Hydrocarb. Radicals (HR)			
	Inorganic (IN)			
	Metal Cation (CN)			
	Inorganic Anion (AN)			
	Abnormal Mass (MASS)			
	General Metal (MF M)			
	Nonstructural (DATA)			
	Registry Number (RN)			

* To simplify reference to a particular query, the user may assign each query a name, which is recorded on the coding form at the top of p. 1. This name will be used in the final encoding of the query and must therefore satisfy the format requirements stated in Sec. 3.5.1.

17

Part III subdivides into three columns. In the first column, a keydesignation is assigned to each key used in the query. A keydesignation is a name invented by the querist that is used to represent a particular key internally in processing a query and externally in error messages. Each keydesignation must begin with a letter and must contain five or fewer letters or digits; special characters (periods, commas, spaces, etc.) are not permitted. The words KEYS and END are also invalid keydesignations, as these have special uses in the query language (see Sec. 3.5.2.2 and Sec. 3.5.3). Otherwise, choice of keydesignations is completely arbitrary, and names significant to the individual user can be readily devised. If the same keydesignation is assigned to two keys in the same query, the first assignment is used. Examples of valid keydesignations are:

CIDS
NOHYD
KEY3
SATD

The following keydesignations are invalid:

TWIGGY   (too many characters)
K.1   (nonalphabetic or non-numeric character)
1CIDS   (begins with a number)
K 1   (imbedded space)

As already discussed (Sec. 2.), search efficiency can be increased by having the keys with the shorter keylists appear first in the input. If keydesignations with some obvious sequence (e.g., K1, K2, K3...) are assigned to the keys in expected order of increasing list length, the optimum ordering of the keys in the input will be immediately evident to the final encoder, even if he has no chemical background.

In the second column the CIDS codes for the keys to be used in the query are listed, one on each line adjacent to the appropriate key designation. In column three, the combination of these keys required to have been assigned to all retrieved compounds is indicated using multipliers and the connectors AND, OR and NOT. (For examples, see Section 4.)

The sections which follow describe the CIDS codes, the assignment, and the uses of each type of molecular formula and structural fragment key. The codes are only briefly summarized; for details of key codes and assignment, see CIDS No. 6 (3).

18

3.2.1 Molecular Formula Keys (CIDS No. 6, pp. 10-11)

CIDS Code

MF el   (Qualitative molecular formula keys)

MF el n   (Quantitative molecular formula keys)

Assignment

The molecular formula (MF) keys are assigned on the basis of the Hill formula. Since each compound has only one such formula, each MF key may be assigned only once to each compound. Demanding that a particular MF key NOT occur eliminates all compounds to which that key has been assigned.

Discussion

The Hill molecular formula represents the total atom content of a compound exclusive of water of hydration. Therefore, compounds which differ in the number and types of addends united to the same parent structure are assigned different sets of molecular formula keys.

Frequently molecular formula requirements can be specified with either an MF statement (Sec. 3.3) or an MF key. In processing a query, the compounds whose locations are obtained from intersecting the keylists must be accessed from the file before the molecular formula statement can be applied. If the key in question is expected to significantly reduce the number of compounds which must be located, use of the key is preferable. Suppose however that a key with a very long keylist (say, MF N 2) is in question, and that the remaining keys in the query severely restrict possible responses. In such a case, the small additional number of potential retrievals that are eliminated by the lengthy MF key does not justify the computer time necessary to test for this key, and simply imposing the MF statement with appropriate restrictions is more efficient.

CIDS does not employ MF keys for the qualitative presence of N or O, since such lengthy keys would be highly inefficient. Requiring that the MF N O or the MF O O keys NOT be assigned, however, eliminates compounds not containing N or O, thus requiring N or O qualitatively in all responses to the other keys in the query.

If a query employs a structural fragment key containing a particular heteroelement, then including the qualitative MF key for that element adds no information,

19

and simply wastes computer time. Suppose, however, that a heteroelement does not occur in any of the structure fragment keys in the query but is contained in a structure to be A/A searched. Since A/A search is a lengthy process, at least the qualitative presence of that element should first be tested for by an MF key or an MF statement in order to fail as many invalid compounds as possible before the A/A search is conducted.

3.2.2 Acyclic-Cyclic Keys (CIDS No. 6, p. 13)

CIDS Code

A-C=n   Where $\underline{n}$ is the number of <u>rings actually drawn</u> in structuring the <u>total compound.</u> The symbol Ph counts as one ring. Rings within a bracketed, subscripted structure are counted only once.

Assignment

Since $\underline{n}$ is the number of rings drawn in structuring the <u>total</u> compound, each compound is assigned only one A-C=n key.

Requiring that a particular A-C=n key NOT be assigned to responses eliminates all compounds structured with <u>exactly</u> $\underline{n}$ rings; compounds structured with more or fewer rings are not eliminated.

Discussion

The value of $\underline{n}$ in each A-C=n key is the total number of rings drawn in structuring the compound, regardless of the number of rings in each of the individual structures that together comprise the total compound. Thus, α-naphthylamine and its hydrochloride,
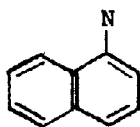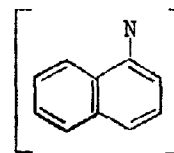
are both structured with two rings are assigned the key A-C=2. However, α-naphthylamine phthalate (2:1)
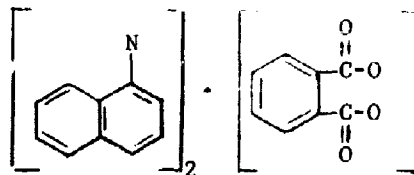
is structured with three rings and is assigned A-C=3. To specify the cyclic character of an individual <u>structure</u> rather than that of a total compound, it

20

is necessary to use the NCN= n keys which tag the number of <u>cyclic nuclei</u> (not necessarily equal to the number of rings) in each structure of every compound (see Sec. 3.2.4).

As the name implies, the acyclic-cyclic keys are used primarily to distinguish cyclic from totally acyclic compounds. The A-C=0 key retrieves compounds having absolutely no rings (total acyclics). Requiring that the A-C=0 key NOT be assigned eliminates all total acyclics from consideration.

3.2.3 Extracyclic Keys (CIDS No. 6, p. 15)

CIDS Codes

$$
\left.\begin{array}{l}
\text{EC1}=n \\
\text{EC2}=n \\
\text{EC3}=n \\
\\
\text{EC4}=n
\end{array}\right\} \text{ where } \underline{n} \text{ is the exact number (0,1,2...) of extracyclic}
\left\{\begin{array}{l}
\text{carbon-carbon double bonds (C=C)} \\
\text{carbon-carbon triple bonds (C}\equiv\text{C)} \\
\text{'Y' carbon configurations } C-\overset{C}{\underset{C}{C}} \text{ or } C=\overset{C}{\underset{C}{(}} \\
\text{'X' carbon configurations } C-\overset{C}{\underset{C}{C}}-C
\end{array}\right.
$$

Assignment

Generally, every structure in a compound is assigned one of each of the four types of extracyclic keys. (Thus, for example, a compound containing two structures is assigned two EC1=$\underline{n}$ keys with the appropriate values of $\underline{n}$ for each structure.) The same EC key can be assigned to more than one structure in each compound, except that the keys EC1=0, EC2=0, EC3=0, and EC4=0 are assigned no more than once to each compound.

Requiring that a particular EC key (e.g., EC2=3) must NOT be assigned to retrievals eliminates all compounds containing a structure to which that key is assigned.

Discussion

The extracyclic keys reflect the degree of unsaturation and branching in a compound. The value of $\underline{n}$ is exact in all cases. Thus compounds having two or three extracyclic C=C bonds are <u>not</u> retrieved by the EC1=1 key.

If a structure is assigned both EC1=0 and EC2=0, its acyclic portions are saturated. Similarly, structures which are assigned both EC3=0 and EC4=0 are unbranched.

21

All of the EC keys having low values of $\underline{n}$ are expected to have very long keylists. Use of these keys can often be avoided by

(1) using wherever possible a specific hydrocarbon radical key (e.g., the isobutyl key HR21E ($C-\overset{\overset{\text{C}}{|}}{C}-C-E1\sim$) instead of an EC3 key)or a specific functional group key (e.g., the allene key FG134 ($\sim C=C=C\sim$) instead of an EC1 key) when a particular fragment is required, and

(2) employing an MF statement with the CONSTRAINTS option to necessitate a required degree of unsaturation.

3.2.4 Number of Cyclic Nuclei Keys (CIDS No. 6, p. 16)

CIDS Code

NCN=n   where $\underline{n}$ is the exact number of cyclic nuclei in a structure

Assignment

Each structure in a compound is assigned one NCN=$\underline{n}$ key with the proper value of $\underline{n}$. The key NCN=0 is assigned no more than once to each compound; the NCN keys with values of $\underline{n}$ greater than zero are assigned as many times as required.

Requiring that an NCN=$\underline{n}$ key NOT occur eliminates all compounds containing one or more structures having exactly $\underline{n}$ nuclei.

Discussion

The value of $\underline{n}$ in each NCN= n key is the total number of cyclic nuclei (not rings) in a single structure. Therefore, the key NCN= 0 retrieves not only totally acyclic compounds, but also all compounds in which an acyclic structure is dot-connected to a cyclic structure.

Since the NCN= n keys tag the cyclic character of each structure, these keys are extremely useful in retrieving a particular structure of known ring content regardless of any additional cyclic or acyclic addends that may be present. For example, the benzoate anion
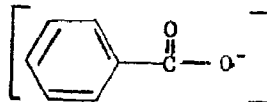
is retrieved by the key NCN= 1 regardless of whether its associated cation is acyclic, such as tetramethylammonium   or cyclic, such as
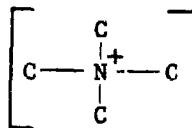
N-methylpyridinium

22
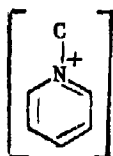
3.2.5 Cyclic Nuclei - Nonhydrogen Attachments Key (CIDS No. 6, p. 17)

CIDS Code

DACN=n   where n is the total number of nonhydrogen attachments to all cyclic nuclei in a single structure.

Assignment

One DACN=n key with the appropriate value of n is assigned to each structure which contains at least one cyclic nucleus. Compounds containing more than one structure can be assigned a particular DACN=n key more than once if required, except that the key DACN=0 is never assigned more than once to each compound.

Requiring that DACN=n NOT occur eliminates from retrieval all compounds containing one or more structures to which that key has been assigned.

Discussion

The value of n in each DACN= n key is the total number of nonhydrogen attachments to all nuclei in a single structure. Attachment may be either to an acyclic atom or to an atom in another nucleus. The multiplicity of the connecting bond is immaterial: a single, double or triple bond each counts as one attachment.

Since the DACN keys are assigned only to structures containing at least one cyclic nucleus, the key DACN= 0 retrieves structures which consist of un-substituted cyclic systems in their various states of hydrogenation. For values of n greater than zero, each structure that responds to a key DACN= n contains a maximum of n cyclic nuclei. Depending on the other requirements specified in a query, the total number of cyclic nuclei in a structure may be implied by the DACN key, thus eliminating the need for an NCN key.

3.2.6 Generic Cyclic Nuclei Keys (CIDS No. 6, pp. 18-21)

CIDS Code	Key name
GCN1=n	Ring Count
GCN2=n, m, ...	Redundant Numerical Ring Population
GCN3=el$_1$ n$_1$ el$_2$ n$_2$ el$_3$ n$_3$ ...	Elementary Ring Population
GCN4=el$_1$ n$_1$ el$_2$ n$_2$ el$_3$ n$_3$ ...	Skeleton Molecular Formula
GCN5=n	Double Bonds in Nucleus
GCN6=n, m, ...	Heteroelement Distribution

Assignment

GCN1 Each nucleus which appears in the CIDS structure of a compound is assigned one GCN=n key where n is the smallest number of smallest rings that account for the entire nucleus.

GCN2 Each nucleus is assigned one GCN2 key having the form GCN2=n for one ring nuclei; GCN2=n,m for two ring nuclei; etc., where n,m, ... give the number of atoms in each of the GCN1 rings.

GCN3 Each ring in a nucleus is assigned a GCN3 key based on the Hill-ordered formula for all of the atoms in the ring. These keys are assigned to (a) any ring that is a member of any smallest set of smallest rings and (b) any ring having 8 or fewer atoms.

GCN4 Each nucleus is assigned a GCN4 key based on the Hill-ordered formula of the complete skeletal framework of the nucleus.

GCN5 Each nucleus in a compound is assigned one GCN5=n key where n is the total number of double bonds in the nucleus.

GCN6 Each one-ring nucleus containing two or more heteroatoms and not more than fifteen atoms in all is assigned one GCN6 key having the form GCN=n,m,... where n,m,... identify the relative positions of heteroatoms around the ring.

Requiring that a particular GCN key is NOT assigned to retrievals eliminates all compounds containing a nucleus to which that key is assigned.

Discussion

The generic cyclic nuclei (GCN) keys tag six types of characteristics of ring systems. These keys enable the user to define a required family of cyclic nuclei by demanding the set of GCN keys that tag the characteric structural features of that family. As the set of keys becomes more restrictive, the family of nuclei that responds becomes smaller, perhaps having as few as one or two members. For maximum search efficiency, only the minimum set of GCN keys that defines the required family of nuclei is stipulated. Thus a search for all $C_{13}N$ nuclei having three six-membered rings would demand the skeleton molecular formula key GCN4= C 13 N 1 and the numerical ring population key GCN2= 6,6,6; the ring count key GCN1= 3 which tags all three ring nuclei is not demanded since all nuclei that respond to GCN2= 6,6,6 contain exactly three rings.

GCN keys are assigned to every nucleus, regardless of whether or not the nucleus is also tagged with one of the specific cyclic nuclei (SCN) keys described in the following section.

24

3.2.7 Specific Cyclic Nuclei Keys (CIDS No. 6, pp. 22-45)

CIDS Code

> SCNn   where n is the one to three digit number for the particular
> nucleus.

Assignment

Each SCN key is assigned as often as the corresponding nucleus appears in the
CIDS structure of the total compound. The SCN key for the benzene nucleus (SCN48)
is assigned both to the fully structured nucleus and to the abbreviated representation Ph.

Specifying that a particular SCN key must NOT be assigned to retrievals
eliminates every compound that contains that nucleus in any of its component
structures.

Discussion

The specific cyclic nuclei keys enable the specification of each of approximately 150 commonly occurring nuclei using only a single key per nucleus. Each
of these nuclei is also assigned all appropriate GCN keys in addition to its own
specific key.

In general, a family of cyclic nuclei is specified by demanding the minimum
set of generic cyclic nuclei (GCN) keys which define that family. However, if
every member of the required family is tagged with an SCN key, then the entire
family is most efficiently specified by simply demanding the SCN key for each
member.

3.2.8 Specific Functional Group Keys (CIDS No. 6, pp. 46-118)

CIDS Code

> FGn   where n is a one to three digit number and the fragment
> is not attached to a ring.

> FGnR   where n is a one to three digit number and the fragment
> has one or more attachments to ring(s).

Assignment

The specific functional group keys are assigned as often as the corresponding

25

fragment appears in the CIDS structured compound. If the fragments corresponding to two keys overlap, both keys are assigned. If the fragments for two keys completely coincide, only the more specific key is assigned. If one key is completely contained in a larger key, only the larger key is assigned unless the larger key is FG23, FG121 or FG122 (or the corresponding FGnR keys), in which case both the smaller and larger keys are assigned.

Requiring a functional group key NOT to be assigned to responses eliminates from retrieval all compounds containing the tagged fragment.

Discussion

To retrieve all compounds containing a particular functional group regardless of whether or not that group is attached to a ring, a query must demand both compounds assigned the acyclic-attached key FGn for that fragment and compounds assigned the ring attached key FGnR. Similarly, to eliminate all compounds containing a particular group, both the FGn and the FGnR keys for that fragment must be negated.

Requiring a key to be assigned n times to each response retrieves all compounds assigned that key n or more times. To retrieve only compounds assigned that key exactly n times, the "n AND NOT n+1" strategy (Sec. 2) must be used. Sometimes, however, restrictions imposed on the molecular formula with either MF keys or the MF statement automatically prohibit multiple occurrences of a particular functional group. For example, a query to retrieve all acyclic compounds in which the only functional group is a single $-NO_2$ might employ the specific functional group key for nitro, FG154, and the quantitative molecular formula key MF N 2. Since all compounds having more than one $-NO_2$ group are failed by the MF N 2 key, the "n AND NOT n+1" strategy need not be used in connection with the nitro key.

3.2.9 Nonspecific Diatomic Functional Group Keys (CIDS No. 6, pp. 119-123)

CIDS Code

NDn   where n is a one or two digit number and the fragment is not attached to a ring.

NDnR   where n is a one or two digit number and the fragment has at least one attachment to a ring.

Assignment

A key in this category is assigned only in the absence of an appropriate specific functional group key. Overlap of these keys with each other and with specific functional group keys is frequently encountered. Rules for assignment of specific functional group keys (Sec. 3.2.8) apply here as well.

Discussion

This category of CIDS keys serves well to reduce the population of specific functional group keys which would otherwise be required in the system. Comments on the specific functional group keys (Sec. 3.2.8) apply here also.

3.2.10 Nonspecific Monatomic Functional Group Keys (CIDS No. 6, pp. 124-125)

CIDS Code

NMn   where $\underline{n}$ is a one or two digit number and the fragment is <u>not</u> attached to a ring.

NMnR   where $\underline{n}$ is a one or two digit number and the fragment is attached to a ring.

Assignment

A key in this category is assigned only in the absence of a more specific functional group key, i.e. either an FG key or an ND key. Rules for assignment of specific functional group keys (Sec. 3.2.8) apply here as well.

Discussion

This least specific category of functional group keys provides assurance that each compound containing a functional group has at least one functional group key assigned to it. Comments on the specific functional group keys (Sec. 3.2.8) apply here as well.

3.2.11 Hydrocarbon Radical Keys (CIDS No. 6, pp. 126-145)

CIDS Code

The codes for these keys consist of the letters HR followed by two to five letters and digits.

27

Assignment

Each hydrocarbon radical (HR) key is assigned as often as the corresponding fragment appears in the CIDS-structured total compound.

Requiring that a particular HR key NOT be assigned eliminates every compound that contains the radical tagged by that key.

Discussion

Unlike the functional group keys, if a structural fragment in a compound corresponds to two or more hydrocarbon radical keys, all appropriate keys are assigned. Thus, the generic hydrocarbon radical key HR29E (C_5-El$\sim$) is assigned to <u>every</u> five-carbon alkyl radical attached to a heteroatom, including the <u>n</u>--pentyl radical which is also tagged with the more specific key HR25E (C-$(C)_4$-El$\sim$). Similarly, the methyl radical in N-methylpiperidine   is assigned both
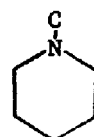
of the equally specific keys HR1E (C-El$\sim$) and HR1R (C-R) because the nitrogen to which this group is attached is both a heteroatom and a ring atom.

Often a family of radicals can be retrieved quite efficiently by imposing proper molecular formula restrictions, thereby eliminating the need for a sizable number of HR keys. For example, a query to retrieve all unsubstituted C_1-C_6 alkanols would require a total of ten hydrocarbon radical keys. This family of alkanols can be retrieved much more efficiently simply by including an MF statement which limits the carbon count to the range 1 to 6, and requires the relationship $H = 2*C + 2$ between the hydrogen count H and the carbon count C to necessitate complete saturation.

28

3.2.12 Inorganic Compound Key (CIDS No. 8, p. 145)

CIDS Code

    IN

Assignment

The IN key tags every inorganic compound, i.e., all non-carbon compounds plus metal carbonates, bicarbonates, cyanides , isocyanides, carbides and carbonyls.

Requiring the IN key NOT to be assigned to responses eliminates all inorganics from consideration.

Discussion

Inorganic compounds are not currently admitted to the CIDS file, although techniques for handling this type of compound have been devised. When the programming to implement these techniques is completed and inorganic compounds are entered into the file, the IN key will enable the querist either to confine his search to inorganics or to eliminate inorganics from consideration.

3.2.13 General Metal Key (CIDS No. 6, p. 145)

CIDS Code

    MF M

Assignment

The general metal key operates as a qualitative molecular formula key, described in Section 3.2.1. For the CIDS definition of "metal", see CIDS No. 6, p. 145.

Discussion

The general metal key is assigned to every compound whose molecular formula contains any number of atoms of any metal(s), regardless of the capacity in which the metal atom is functioning. If the user is interested

29

specifically in metal atoms functioning as bare metal cations, the general
metal cation key discussed below is used.

3.2.14 General Metal Cation Key (CIDS No. 6, p. 145)

CIDS Code

CN

Assignment

The CN key is assigned qualitatively to each compound that contains a bare
metal cation or the ammonium (NH_4^+) cation. It is assigned no more than once to
each compound, even if the compound actually contains more than one such cation.

Specifying that the CN key does NOT occur eliminates from retrieval all
compounds which are structured with a metal or ammonium cation.

Discussion

The CN key distinguishes organic acids from their normal and acid metal
or ammonium salts. Especial attention is called to the fact that organometal
cations, e.g., are <u>not</u> assigned this key.

Ammonium salts exclusively can be retrieved by employing the CN key
to obtain potential responses having a metal or ammonium cation and then
eliminating those compounds that contain any metal by negating the general
metal key MF M (Sec. 3.2.13).

3.2.15 General Inorganic Anion Key (CIDS No. 6, p. 145)

CIDS Code

AN

Assignment

The AN key is assigned qualitatively, i.e., it may be assigned only
once to each compound even if the compound contains more than one inorganic
anion.

Specifying that the AN key NOT occur eliminates all compounds containing
any number of inorganic anions.

30

Discussion

The AN key is used in retrieving the various "ium" salts (e.g., onium, inium, ylium compounds) in which the cation of an organic base is associated with an inorganic anion.

3.2.16 Abnormal Mass (Isotope) Key (CIDS No. 6, p. 146)

CIDS Code

MASS

Assignment

The MASS key is qualitatively assigned to every compound containing any number of specified isotopes of any element(s).

Specifying that the isotope must NOT be assigned to retrievals eliminates all compounds that contain any specified nuclide.

Discussion

The isotope key MASS tags every compound that contains any specified nuclide(s), including deuterium (D) and tritium (T). Demanding the MASS key therefore retrieves all isotopically tagged compounds, regardless of the total number of isotopes present or of the element type, mass number or location of each in the total molecule.

To specify either the element type or the element type and mass number of a tagged atom and to require the atom to participate in a particular functional group, an A/A search is required. Whenever an A/A search is conducted for a specified isotope, the MASS key should still be demanded in order to minimize the number of compounds which must be accessed from the file and searched.

3.2.17 Nonstructural Information and Data Keys

CIDS Code

DATA cc   where cc is the two-letter code for one of the nonstructural
             descriptors listed below.

31

Assignment

A compound is assigned the corresponding DATA key for each two-letter descriptor code that is associated with one or more references in the compound record.

Discussion

The molecular formula and structural fragment keys described in previous sections and the FORMULA statement and atom-by-atom search discussed later enable effective search of compounds currently admitted to the CIDS file on the basis of their structural and/or molecular formula characteristics. Techniques for structure and formula searches of most currently inadmissible types have also been devised, although these have not yet been implemented. However, the enormous area of chemical knowledge beyond structure and formula, including applications, reactions and the various physical constants, has yet to be organized for efficient computer search.

The technique for handling nonstructural material that is currently being explored in CIDS utilizes an open-ended list of nonstructural descriptors, each of which is represented by a two-letter mnemonic code. In compiling a compound record for admission to the file, the chemical editor associates with each reference the set of two-letter descriptor codes that identifies the types of nonstructural information discussed in that reference. When the record is processed by the CIDS file-building programs, the compound is assigned each DATA key that corresponds to a descriptor code appearing in one or more of the references.

In its current experimental mode, approximately 60 descriptors are employed. They have been compiled by Edgewood Arsenal based on experience with the types of nonstructural information commonly encountered in references, and taking into consideration expected user needs. The current set of descriptors with their codes is as follows:

Descriptor	Code
Applications	AP
Activity Coefficient	AC
Analytical Detection	AD
Analytical Determination	AN

(continued).

32

Descriptor	Code
Industrial Applications	BA
Boiling Point	BP
Biological Suppressant	BS
Crystalline Form	CF
Chromatographic Methods	CM
Cost	CO
Critical Pressure	CP
Color	CR
Critical Temperature	CT
Dyestuff Application	DA
Dissociation Constants	DC
Derivatives	DV
Entrophy	EN
Electron Spin Resonance Spectrum	ES
Free Engry	FE
Geometric Isomers	GI
Heat Capacity	HC
Heat of Dilution	HD
Heat of Formation	HF
Heat of Solution	HS
Heat of Sublimation	HU
Heat of Vaporization	HV
Hydrates	HY
Synthesis Intermediate or Starting Product	IA
Ionization Constants (pKa, pKb)	IC
Incapacitating Dose (Dosage)	ID
Eye Irritant or Lacrimator	IE
Infrared Spectrum	IR
Kinetics of Hydrolysis	KH
LD_{50} (Dosage)	LD
(Med) Minimum Effective Dose (Dosage)	ME
Melting Point	MP
Mass Spectrum	MS
Nuclear Magnetic Resonance Spectrum	NS
Optical Isomers	OI
Optical Rotation	OR
Polarography	PO
Purification	PU
Respiratory Inhibition	RE
Refractive Index	RI
Raman Spectrum	RS
Stability	SB
Solvent of Crystallization	SC
Specific Gravity	SG
Specific Heat	SH
Hammett Sigma Value	SI
Solubility	SO
Specifications	SP
Surface Tension	ST
Suppliers	SU

(continued)

Descriptors	Code
Solvates	SV
Synthesis	SY
Triple Point	TP
Ultra Violet Spectrum	UV
Viscosity	VI
Vapor Pressure	VP

3.2.18 Registry Number Keys

CIDS Code

RN 000ddd   where ddd are the three low order digits of the CIDS master registry number of the compound.

RN dddXXX   where ddd are the three higher-order digits of the CIDS master registry number of the compound.

Assignment

Each compound registered in the CIDS file is assigned two registry number (RN) keys, one tagging the three low-order digits of the master registry number and the other tagging the next three higher order digits.

Discussion

Each compound registered in the CIDS file is assigned a master registry number (MRN) consisting of the letter "A" followed by seven digits. Users of the system have found that it would be useful to be able to access compounds from the search file on the basis of this registry number alone. This capability might have been accomplished by assigning a unique registry number key to each compound in the file. However, since the system is limited by core size as to the total number of different keys that can be accommodated, this technique is unacceptable for any sizable file of compounds.

A solution was found to be the assignment of two keys to each compound, one designating the three low-order digits of the registry number and the other designating the next three higher order digits. (Thus, the compound whose registry number is A0104286 is assigned the two keys RN 000286 and RN 104XXX.) Since the size of the CIDS file has required use of only six digits in the MRN, each pair of one high with one low-order RN key corresponds to exactly one compound. Plans have been made to enable specification in a

34

query of the complete registry number by means of a single key that will be expanded into two keys internally.

The high-order digits RN keys serve as a means of separating the file into blocks of 1000 compounds. This is a useful capability, because the search system is designed so that only the first 1000 responses to the keys become candidates for the formula statement and the A/A search, and for eventual output. A query in which the keys obtain more than 1000 responses can be broken down into several subqueries, each of which uses the same original set of keys plus a high-order digits RN key. Each subquery is thereby restricted to a different block of 1000 compounds, and a manageable number of responses to the keys in each case is insured.

3.3 Part IV: Molecular Formula Statement (Optional)

Part IV of the coding form (illustrated below) contains the user's specification of the restrictions to be imposed by the Molecular Formula (MF) Statement on all compounds that respond to the keys. The actual encoding of this data for input will be performed in Part II as described in Section 3.5.2.3.

Illustration: Query Coding Form, Part IV

IV.	MOLECULAR FORMULA STATEMENT						
	Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

The molecular formula information contained in the CIDS file can be summarized as follows:

1. Every compound record contains a Hill molecular formula representing the total atom content of the compound exclusive of water of hydration.

2. If the anhydrous portion of the compound is a type structured in CIDS as

35

a combination of two or more distinct smaller compounds*, the total formula is also expressed as the dot-connected formulas of each of the smaller compounds preceded by appropriate multipliers. Each of these dot-connected formulas is referred to as an addend molecular formula.

3. If the compound is a hydrate, both the Hill formula and (when present) the string of addend formulas are dot-connected to nH_2O, where $\underline{n}$ is the total number of water molecules. ($\underline{n}$ is an integer or a proper or improper fraction.)

The MF statement is used to impose restrictions on the Hill and/or up to four organic or inorganic addend molecular formulas†, and to require that the molecular formula contain water. Only one formula statement can be used in each query‡. A compound can be retrieved only if its molecular formula satisfies all of the conditions set down in the formula statement. The restrictions imposed on each formula are specified in Part IV of the coding form as follows:

The column headed:	Is completed as follows:
Formula type	If restrictions are to be imposed on the Hill formula, the first "Formula type" block is marked HILL, and the restrictions are stated in the adjacent columns. If no restrictions are being placed on the Hill formula, the "Formula type" block is marked ADDEND, and the restrictions contained in the adjacent columns are imposed on an addend molecular formula. The second and all subsequent sets of restrictions are assumed to apply to addend molecular formulas, so no additional formula types need be specified.
RESTRICTED	If the set of elements listed under Element type (see below) is the exact set contained in the formula, the RESTRICTED box is marked yes; otherwise, it is left blank.
Element type	Symbols of the elements whose presence is to be tested for in the formula statement are listed in Hill order; in addition to the standard element symbols, including D (deuterium), T (tritium) and the individual halogens, the symbol X can be used to represent any halogen atom (F, Cl, Br, or I).

---

* The rules for structuring compounds admitted to the CIDS file are contained in Chemical Editing Conventions I (5).

† Part IV provides space for specifying the restrictions on two formulas; space for restrictions on additional formulas may be improvised as required.

‡ If more than one formula statement is input in a single query, the last one typed in is used.

36

Exact count	Adjacent to each element symbol is specified in the
Lower bound	appropriate column the exact count, or a lower and/or
Upper bound	upper bound for each element*. The values 0,1,2, ...
	are valid. If these three columns are left blank, only
	the qualitative presence of that element is required.

CONSTRAINTS — Algebraic relationships between two element counts of the form

$$el = a*el \pm b$$

a = 1 to 63, b = 0 to 31,
el = the standard symbol (or the symbol X) of the element whose count is being compared

are listed in the CONSTRAINTS column. If more than one pair of elements in a given formula are appropriately related, multiple relationships can be specified. When $a=1$, the a and the * can be omitted; when $b=0$, the term $\pm b$ can be omitted.

The presence of water in the molecular formula is specified by demanding that responses have an addend molecular formula which is RESTRICTED to exactly two hydrogens and exactly one oxygen.

For examples of specifying molecular formula statements, see Section 3.5.2.3 or Section 4.

3.3.1 Use of the Molecular Formula Statement

The objective in efficient query formulation is to specify the **minimum** set of characteristics necessary to retrieve the desired compounds. Frequently data specified in the formula statement **eliminates** the need for a number of structural fragment keys. Suppose, for example, all tertiary amines of the form

$$CH_3-N-R$$
$$\quad\quad |$$
$$\quad\quad CH_3$$

where R is any C_1-C_7 alkyl are required. A total of eleven specific and generic hydrocarbon radical keys are required to specify all of the alkyl groups that could occur in such compounds. However, since all members of this family have the general molecular formula $C_nH_{2n+3}N$, simply imposing a RESTRICTED formula statement with the constraints $H = 2*C + 3$ (along with the tertiary amine key and two methyl keys) assures that all respondents have only the permitted alkyl substitutions.

---

* The count that is specified for the symbol X refers to the total number of halogens in the formula. Thus, for example, demanding "exactly three X" retrieves compounds containing three atoms of Cl, or two Cl and one Br, etc.

On the other hand, the structural fragment keys employed may imply certain molecular formula characteristics, making explicit specification of these characteristics in the formula statement unnecessary. Consider the following examples.

(1) A querist who wishes to retrieve a family of disulfides is aware that all true responses contain two or more S atoms. However, all compounds that respond to the structural fragment key for disulfides ($\sim$S-S$\sim$) must contain a minimum of two S, thus making it unnecessary to stipulate a lower bound of two for S in the formula statement.

(2) Acyclic aldehydes from C_3 to C_{10} containing one double bond and no other functional groups are to be retrieved. Keys are used to require exactly one aldehyde group, exactly one $\sim$C=C$\sim$, exactly one O and no cyclic nuclei. The specifications required of the Hill formula in the formula statement are

(a) RESTRICTED

Since all true answers can contain only the elements C, H and O, the RESTRICTED option is exercised.

(b) CONSTRAINTS H = 2*C — 2

All true responses have the general formula $C_nH_{2n-2}O$; thus the relationship H = 2*C — 2 can be required.

(c) Element Counts

Carbon: All true responses must contain 3-10 C atoms. However all responses to the keys $\sim$C=C$\sim$ and ---C=O contain at least 3 carbons. Therefore only the upper bound of 10 C atoms must be stipulated.

Hydrogen: The H count has already been expressed in the CONSTRAINTS as a function of the count of 3 to 10 C atoms. Since the formula is RESTRICTED, hydrogen must be demanded, but only qualitatively.

Oxygen: Since the formula is RESTRICTED, oxygen must be listed; however, the keys have already demanded an exact count of one oxygen, so no count need be specified in the formula statement.

It is evident then that molecular formula requirements that are implicit in the structural fragment keys should not be repeated in the formula statement. However, this is not true of formula requirements implicit in structures to be tested

38

for by A/A search. In actual operation, only the compounds that satisfy the keys and formula statement are tested by the A/A search. Since A/A search is such a lengthy process, the keys and the formula statement should be used to fail as many compounds as possible before the A/A search is begun. Therefore if a structure to be A/A searched contains a heteroelement whose presence is not already necessitated by the structural fragment keys, at least the qualitative presence of that element should first be established either with a molecular formula key or in the formula statement.

3.4 Part V : Atom-by-Atom Search (Optional)

Structural fragments to be located by atom-by-atom (A/A) search are constructed and encoded in Part V (illustrated below). Each encoded structure is assigned a name used to represent the structure when the query is processed, and the number of times each structure is required to occur in each retrieval is specified

Illustration: Query Coding Form, Part V

V.

ATOM-BY-ATOM SEARCH		
Structure(s):		
Structure name	Structurespecification	No. of Occurrences

3.4.1 Use of the Atom-by-Atom Search

Identification by computer of a particular structural fragment in the record of a compound in the CIDS file is a time-consuming process. If retrieval of structural families required that the record of every compound on file be examined each time a query was input, the system would at best be highly uneconomical, and real time operation would be impractical. To avoid the need for examining individual compound records at search time, CIDS identifies and tags with keys significant structural features of each compound before the compound is stored in the file. Thus when a search is initiated, the required structural family can usually be obtained by combining the appropriate structural fragment keys, supplemented by the molecular formula keys and/or the formula statement as necessary.

Sometimes the keys and formula statement alone will not retrieve exclusively the desired structural family. If the percent of false responses is expected to be low, final discrimination is most economically accomplished by simply examining the output visually. However, when both the number and percent of false drops is expected to be high, it is possible to have the record of each potential response atom-by-atom searched for the precise structural feature(s) which characterize true responses. Examples of situations in which an A/A search might be profitable are:

(1) to require a particular juxtaposition of two (or more) functional groups. (Examples: aminoacids in which the $-NH_2$ is beta to the carboxyl; benzenesulfonamides having an amino group para to the sulfonamide (i.e., sulfanilamides); etc.)

(2) to require that a particular atom carry a charge, with or without specifying the exact value of the charge;

(3) to require a specified isotope of a certain element to be contained in a particular functional group, with or without specifying the exact mass number of the isotope;

(4) to demand a structural fragment in which one or more atoms may be any of a specified set of elements. (Example: Dialkyl compounds of group

40

IIB metals, i.e., compounds containing the structure* ~C-E2-C~ where E2 is zinc, cadmium or mercury.)

When it is necessary to conduct an A/A search, only the <u>minimum</u> fragment which will detect the required structural feature should be specified. For example, consider a query to retrieve all monocyclic compounds in which two halogens are substituted on the same carbon atom of a cyclohexane nucleus (other substitutions are permitted). While the keys can assure that cyclohexane is the only nucleus, and that the ring has at least two halogen substitutions, A/A search is necessary to require both halogens to be attached to the same atom. The complete structure that characterizes true responses is*
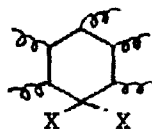
However, since cyclohexane is the only nucleus present, only the fragment
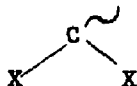
need be specified, with suitable indication that the carbon atom is a ring member.

3.4.2 Structuring the Fragment

The structural fragment must be constructed from the atom symbols and bond types described in Tables I and II. Charges and specified atomic masses are indicated in their usual locations, i.e., to the above right and above left of the atom symbol respectively.

If the number of bonds drawn to an atom is less than the assumed CIDS valence for that atom (four for carbon; otherwise, the lowest common valence), the remaining valence is assumed to be satisfied by hydrogen. Therefore, the only hydrogen atoms that must be shown explicitly are those attached to an atom whose current valence is greater than its assumed CIDS valence.

* For explanations of the symbols, see Tables I and II which follow.

41

TABLE I. ATOM SYMBOLS USED IN ATOM-BY-ATOM SEARCH

The following atom symbols are used in constructing query fragments:

Symbol	Meaning
Any of the standard international symbols for the elements	One atom of the element.
E*l*	One atom of any heteroelement, i.e., any element except C or H.
EE	One atom of any element except H.
D	One atom of deuterium.
T	One atom of tritium.
X	One atom of any of the halogens F, Cl, Br or I.
E1 E2 E3 E4	Each of these symbols is available to be defined by the user to mean either (a) one atom of any element in a list of up to eight element types selected by the user, or (b) one atom of any element except those appearing in a list of up to eight element types selected by the user.

TABLE II. BOND SYMBOLS USED IN ATOM-BY-ATOM SEARCH

The following bond types are used in constructing query fragments:

Bond Type	Hand Representation	Meaning	
		As interior* bond	As hanging* bond
1	——	Single bond	Single bond to an unshown carbon atom
2	═══	Double bond	Double bond to an unshown carbon atom
3	≡≡≡	Triple bond	Triple bond to an unshown carbon atom
4	—⊖—	Resonant ring bond	Resonant ring bond to an unshown carbon atom
5	----	No meaning	Single bond to an unshown atom of C or H
6	∿	Any type bond (a "don't care" interior bond)	No meaning
7	⌒⌒⌒	No meaning	Any number or type of non-ring bond(s) to any unshown atom(s) (an acyclic "don't care" hanging bond)
8	∿	No meaning	Any number or type(s) of ring or non-ring bond(s) to any unshown atom(s) (a "don't care" hanging bond)

*An interior bond connects two atoms within a structural fragment. A hanging bond connects an atom in a fragment to an unshown atom, i.e., an atom that is not part of the fragment.

The following points regarding the construction of fragments are worthy of special note:

(1) <u>Ring and Non-Ring Members</u>   The A/A search distinguishes between atoms and bonds which must be members of rings, those which must not be members of rings and those which may or may not be members of rings. Therefore the ring, non-ring or "don't care" character of each atom and bond in the fragment must be clearly evident.

(2) <u>Use of the Halogen Symbol X</u>   If the symbol X (representing one atom of any of the halogens F, Cl, Br or I) appears more than once in a single fragment, the atoms in the compound record to which the X's correspond are <u>not</u> <u>necessarily</u> atoms of the same halogen.

(3) <u>Defining the Symbols E1-E4</u>   If any of the symbols E1 through E4 are used, the user must provide for each either (a) a list of up to eight element types with which that symbol must correspond, or (b) a list of up to eight element symbols with which that symbol must not correspond. If one of these four symbols is used in more than one fragment in a single query, the symbol must be defined <u>in each fragment</u>.

(4) <u>Retrieval of Isotopes</u>

(a) An atom in a query structure for which no mass is specified can correspond to any isotope (i.e., normal or "abnormal" mass) of that element in the compound record.

<u>Example</u>: An A/A search for compounds containing the fragment
-C=N-O---   retrieves   C-C$\underset{}{\overset{15}{=\!=\!=}}$N-O   and   C-C$=$N-O as well as C-C=N-O.
                                                                                                      D

(b) An A/A search for a particular specified isotope retrieves compounds in which the appropriate atom has the <u>exact</u> mass number specified.

<u>Example</u>: An A/A search for compounds containing the fragment

$$
\begin{array}{c}
O \\
\| \\
---\overset{33}{=\!=\!=}P-O--- \\
| \\
O---
\end{array}
$$

does <u>not</u> retrieve the compound

$$
\begin{array}{c}
O \\
\| \\
C-C\overset{32}{=\!=\!=}P-O \\
| \\
O
\end{array}
$$
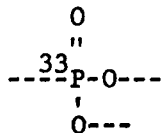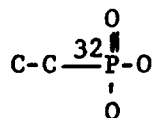
due to the different mass numbers of the phosphorus atoms.

44

(c) Compounds containing an atom permitted to be <u>any</u> specified isotope of a particular element are retrieved by an A/A search in which the atom in question has as its mass number the indefinite value $\underline{n}$.

<u>Example</u>: An A/A search for the fragment

$$---\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O---}{|}}{\overset{\displaystyle ||}{n\text{-}P\text{-}O--}}}$$

retrieves both $C\text{-}C\xrightarrow{}\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O}{|}}{\overset{\displaystyle ||}{{}^{33}P\text{-}O}}}$ and $C\text{-}C\xrightarrow{}\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O}{|}}{\overset{\displaystyle ||}{{}^{32}P\text{-}O}}}$.
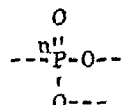
It does not retrieve $C\text{-}C\text{--}\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O}{|}}{\overset{\displaystyle ||}{P}}}\text{--}O$ because the mass number of the

phosphorus atom is not specified.

(5) <u>Retrieval of Charged Compounds</u>

(a) An uncharged atom in a query structure can correspond to any uncharged or negatively charged atom of the same element in the compound record. It can correspond to a positively charged atom in the record only if the atom in the query structure has a "don't care" bond (type 6, 7 or 8) attached to it.

<u>Example</u>: An A/A search for the fragment

$$---\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O}{|}}{\overset{\displaystyle ||}{As\text{-}S}}}$$
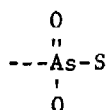
retrieves the compound $\left[C\text{-}C\text{-}\overset{\displaystyle O}{\underset{\displaystyle \underset{\displaystyle O}{|}}{\overset{\displaystyle ||}{As}}}\text{-}S^{-} \right] \cdot K^{+}$, even though the sulfur atom in

the fragment does not have a charge. However an A/A search for the

fragment $-\overset{|}{\underset{|}{P}}-$ does <u>not</u> retrieve the compound $\left[C\text{-}\overset{\displaystyle C}{\underset{\displaystyle C}{\overset{\displaystyle |}{\underset{\displaystyle |}{P^{+}}}}}\text{--}C \right] \cdot Cl^{-}$

because the phosphorus atom in the fragment does not have either a positive charge or a "don't care" bond.

(b) To require an atom to be either positively charged or negatively charged without specifying the exact value of the charge, an A/A search is conducted in which the appropriate atom has either the indefinite positive charge +n or the indefinite negative charge -n.

Example: An A/A search for the fragment — Sn^{+n}

retrieves the following compounds:

$$\left[C-C-Sn^+ \right] \cdot Cl^-$$

$$\left[\begin{array}{c} C \\ | \\ C \\ | \\ C-C-Sn^{++} \end{array} \right] \cdot 2Cl^-$$

$$\left[\begin{array}{c} C-C \\ | \\ C-C-Sn^+-C-C \end{array} \right] \cdot Cl^-$$
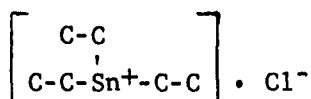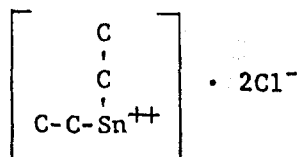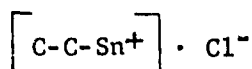
(6) **Retrieval of Complete Structures** An A/A search can be performed for a fragment having no hanging bonds, i.e., for a complete chemical structure. The structure may or may not be charged.

Example: A query to retrieve all salts of trichloroacetic acid might include an A/A search for the complete trichloroacetate anion,

$$Cl \underset{\underset{Cl}{|}}{\overset{\overset{Cl}{|}}{\underset{}{C}}} \overset{\overset{O}{\parallel}}{C}-O^-$$
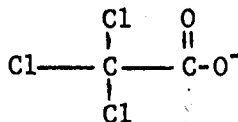
(7) **Retrieval of Deuterium and/or Tritium Compounds** If an atom in a query fragment is required to be deuterated (tritiated)*, the user must either (a) specify the exact number of deuteriums (tritiums) attached to the atom, or (b) specify that the atom must be deuterated (tritiated) but that the exact number of D (T) attachments is unrestricted.

3.4.3 **Structure Name and Number of Occurrences**

Each structural fragment is assigned a name which is used to represent that fragment internally in processing and externally in error messages. Rules for constructing structure names are the same as for keydesignations (Sec. 3.2), namely, each name must begin with a letter and contain five or fewer letters

---

* A deuterated (tritiated) atom is one to which deuterium (tritium) is bonded.

and digits, with no special characters (periods, spaces, etc.) permitted. Since the words KEYS and END have other uses in the query language, these are not proper structure names. If two structures in the same query are assigned the same name, the first assignment is used. A structure may be assigned the same name as a key in the same query.

The number of times each fragment is required to occur in each response is specified as with the keys, using multipliers (1,2,3...), the connectors AND, OR and NOT and parentheses as required. Since any or all of the fragments tested for can be negated, the A/A search can be used to eliminate a structural family as well as to retrieve a family.

3.4.4 Structurespecification

A structurespecification is an encoded linear representation of the chemical structure or structural fragment to be searched for. In processing a query, this encoded structure is compared with the structures of the compounds which have satisfied the keys and the molecular formula statement. Compounds for which a correspondence is found are retrieved or discarded as required by the original question.

The first step in writing a structurespecification for a fragment is to number each atom other than H, D and T consecutively beginning with the number 1. (A maximum of 50 atoms per fragment other than H, D and T is permitted.) This numbering establishes the sequence in which the atoms will be compared with the atoms in the file compound. Although no particular order of numbering is demanded by the system, invalid compounds will be failed most rapidly if the number 1 is assigned to the atom least likely to occur in potential responses, number 2 is the most unusual atom attached to atom 1, 3 is the most unusual atom attached to 1 or 2, etc.

The general form of a structurespecification is

$$\text{atomstring}_1 \quad \text{atomstring}_2 \quad \ldots \quad \text{atomstring}_n \quad (\text{abnormalitystring})$$

where <u>atomstring</u>$_1$ identifies the element type of atom number 1, the atoms attached to 1, and the bond types by which they are attached. (Thus, the structurespecification for a fragment with n atoms other than H, D and T contains n atomstrings.)

<u>abnormalitystring</u> identifies deuterated and tritiated atoms, charged

atoms, and specified isotopes; stipulates when necessary the required valence of a variable valence atom; and lists the elements represented by each of the symbols E1 through E4 if any of these symbols are used in the fragment.

Although the general form for a structurespecification provided above orders the atomstrings by atom number i, any permutation of the atomstrings is permitted. For example, the structurespecification for a three atom fragment could be written "atomstring$_3$ atomstring$_1$ atomstring$_2$" as well as in atom number order, "atomstring$_1$ atomstring$_2$ atomstring$_3$". Specification of attachment in the atomstrings need not be redundant; thus, if the bond between atoms 2 and 3 is cited in atomstring$_3$ it need not also be cited in atomstring$_2$.

The rules for writing atomstrings and abnormalitystrings are provided in the following sections.

3.4.4.1 Writing Atomstrings

A structurespecification contains an atomstring for each atom in the fragment other than H, D and T. In writing atomstrings, attachments to H, D and T are ignored. Each atomstring consists of the following string of characters in the order described:

(1) the atom number: the number i assigned to the atom as previously described.

(2) an asterisk (*) if atom i must be in a ring,

   or

   an apostrophe (') if atom i may or may not be in a ring (i.e., don't care),

   or

   no character. If no * or ' appears, it is assumed that atom i must not be in a ring.

(3) the element symbol of atom i: includes any of the symbols contained in Table I except H, D and T.

(4) a bond description for each bond to atom i that is not cited in another atomstring. Descriptions of bonds (to atom i) that are cited in other atomstrings may also be included, but these are not required. Each bond description consists of the following string of characters:

   (a) the bond type: a one digit number found in Table II.

   (b) an asterisk (*) if the bond must be in a ring,

48

or

an <u>apostrophe</u> (') if the bond may or may not be in a
ring (don't care)

or

<u>no character</u>. If no * or ' appears, it is assumed
that the bond must not be in a ring.

(c) the <u>attached atom number</u>: the number assigned to the
atom attached to atom $\underline{i}$ by this bond. The at-
tached atom number for all hanging bonds is O.

(d) a <u>hyphen</u> (-) if another bond description for atom $\underline{i}$
immediately follows,

or

a <u>period</u> (.) if the last bond description in atomstring$_i$
has been cited.

In summary, each atomstring has the following general form:

atom# ring   el   bond ring attchd——bond ring attchd ... —— bond ring attchd
    char.   type char. atom#   type char. atom#   type char. atom#

in which $\underline{el}$ is a one or two letter element symbol, and the ring character is
either an apostrophe, an asterisk or no character at all. If no bond descrip-
tions are included in an atomstring, the period immediately follows the element
symbol.

<u>Example</u>   Consider an A/A search for compounds containing the fragment

     ---C≡══Si∿   in which neither atom may be in a ring.
       2   1
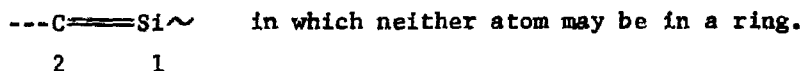
The structurespecification for this fragment is written as follows:

(1) Number the atoms. Since silicon is the more unusual atom, it is
assigned the number 1.

(2) Write the atomstrings.

The fragment contains two atoms that are not H, D or T; therefore its
structurespecification contains two atomstrings. Since the bond between these
two atoms may be cited in either or both atomstrings and since there is no
prescribed ordering of the bond descriptions, there is more than one valid
pair of atomstrings for this fragment. One example is

       1Si22-80.   and   2C50.

which are interpreted as follows:

```
Atom number 1
   \   is a non-ring (since no * or ') atom
    \     of silicon
     \      |   attached by a double (type 2)
      \     |      non-ring (since no * or ') bond
       \    |     /    to atom number 2
        \   |    /    /    and
         \  |   /    /    by any ring or non-ring bond(s) (type 8)
          \ |  /    /        to any atoms not displayed in this fragment
           \| /    /          and to nothing else.  (The type 8 bond accom-
            |/    /                modates attachments to H, D and T as
            |    /                 well as attachments to this atom cited
            |   /                  in other atomstrings.)
        1 Si22 - 80.
```

```
Atom number 2
   \   is a non-ring (since no * or ')
    \     carbon atom
     \      attached by a single
      \        non-ring (since no * or ') bond to an atom of C or H
       \      /(type 5 bond)
        \    / outside this fragment
         \  /     and to nothing else except perhaps H, D, T or an
          \/          attachment cited in another atomstring.
        2C50.
```

(3)   The structurespecification for this fragment consists of the two
atomstrings side by side, i.e.,

<div align="center">1Si22-80.   2C50.</div>

3.4.4.2   Writing Abnormalitystrings

If a structure contains a charged atom, an atom of D, T, or any other
"abnormal" isotope, or any of the symbols El-E4, its structurespecification
must contain an abnormalitystring. An abnormalitystring may also be required
if the structure contains an atom whose valence varies in different compounds.
In writing a structurespecification, the abnormalitystring is enclosed in
parenthesis and follows the last atomstring.

The general form of the abnormalitystring is:

<div align="center">term   term ... term</div>

term   may be any of the six different types of terms described below. An
       abnormalitystring may contain as many terms of each type as are required
       to fully describe the structure, and the ordering of the terms within
       the parentheses is completely arbitrary.

<div align="center">50</div>

a. <u>Charge abnormality</u>   To require an atom to carry a charge, the abnormalitystring must contain a term of the form

$$Ci = v.$$

where i is the number assigned to the atom which carries a charge $v = \pm 1, \pm 2,$ ... To require atom i to be positively charged without specifying the exact magnitude of the charge, v is set equal to the indefinite positive value "+n;" similarly, to require atom i to carry an unspecified negative charge, v is set equal to the indefinite negative value "-n."

b. <u>Attachment to D (deuterium)</u>   If an atom in a fragment is attached to one or more atoms of deuterium, the abnormalitystring must contain a term of the form

$$Di = v.$$

where i is the number assigned to the atom to which v atoms of deuterium are attached, $v = 1, 2, ...$ or the indefinite value n. Setting $v = n$ results in a term of the form $Di = n.$ which requires atom i to be deuterated without specifying the exact number of deuteriums.

c. <u>Attachment to T (tritium)</u>   If an atom in a fragment is attached to one or more atoms of tritium, the abnormalitystring must contain a term of the form

$$Ti = v.$$

where i is the number assigned to the atom to which v atoms of tritium are attached, $v = 1, 2, ...$ or the indefinite value n. Setting $v = n$ results in a term of the form $Ti = n.$ which requires atom i to be tritiated without specifying the exact number of tritiums.

d. <u>Mass abnormality</u>   If an atom in a fragment has an abnormal mass specified above and to the left of the element symbol, the abnormalitystring must contain a term of the form

$$Mi = v.$$

where i is the number assigned to the atom whose mass number is $v = 1, 2,$ ... or the indefinite value n. Setting $v = n$ results in a term of the form $Mi = n.$ which requires atom i to be any specified isotope of the appropriate element, regardless of the exact mass specified.

e. Valence abnormality   Suppose a fragment contains an element whose valence differs in different compounds. (Nitrogen, for example, may have a valence of either 3 or 5.) Suppose further that only compounds in which that element functions with a single, specified valence are to be retrieved. If the required valence is not necessitated by the bonds explicitly cited in the fragment, that valence may be demanded by including in the abnormalitystring a term of the form

$$Vi = v.$$

where i is the number assigned to the atom whose valence must be v ($v = 1, 2$ ...). Note that no sign is associated with the value of v. For additional explanation of the valence abnormality, see the first example below.

f. Defining the Symbols E1, E2, E3, E4   For each of the symbols E1 through E4 that is used in the fragment, the abnormalitystring must contain either a term identifying which elements are represented by the symbol or a term identifying which elements are not represented by the symbol as follows:

A term of the form

$$Ei = el_1, el_2, ...el_n. \qquad (n \leq 8)$$

where Ei is either E1, E2, E3 or E4 and the el's are standard element symbols means that any atom in the fragment represented by the symbol Ei can be any of the elements equated to that symbol. Similarly, a term of the form

$$Ei = NOT\ el_1, el_2, ...el_n. \quad (n \leq 8)$$

(Ei and el's are as described above) means that an atom represented by Ei may be any element except those equated to that symbol.

The el's must be standard element symbols; thus the symbols D, T, X, EE and Eℓ, as well as E1 through E4 themselves, are not valid.

Since four symbols (E1, E2, E3 and E4) are available to be defined in each fragment, a maximum of four of these defining terms can be present in an abnormalitystring. If the same symbol is used in more than one fragment in a single query, that symbol must be defined in the abnormalitystring of each fragment in which it appears. The definitions need not be the same for all of the fragments.

Example 1

Both of the elements N and P can function with a valence of either 3 or 5. In the fragment

$$-\overset{|}{\underset{1}{P}}\overset{+}{=}\underset{2}{N}\sim$$
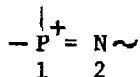
the phosphorus atom is necessarily pentavalent; however, since the wiggly bond represents any number of any type bonds, the N atom can be either trivalent or pentavalent. To retrieve only those members of this family in which the N is, for example, trivalent, the abnormalitystring must contain a term (since N is assigned the number 2)

$$V2=3.$$

The abnormalitystring must also contain a term C1=+1. to require a charge of +1 on the P atom. The complete structurespecification for the above fragment is

$$1P22-10-10. \; 2N21-80. \; (C1=+1. \; V2=3.)$$

Eliminating redundant bond descriptions, this structurespecification can be written

$$1P22-10-10. \; 2N80. \; (C1=+1. \; V2=3.)$$

Example 2

To retrieve all compounds containing a trideuterated carbon atom, an A/A search is performed for the fragment

$$\begin{array}{c} D \\ | \\ D-C\sim \\ | \\ D \end{array}$$
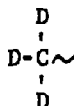
Since atoms of H, D and T are not numbered, the carbon becomes atom number 1. To require three atoms of deuterium to be attached to atom 1, the abnormalitystring must include the term

$$D1=3.$$

The structurespecification for this fragment is

$$1C80. \; (D1=3.)$$

Example 3

An A/A search to retrieve all deuterated heteroatoms other than oxygen demands that responses contain the fragment

~E4   (one or more deuteriums attached to E4)

in which the symbol E4 represents an atom of any element except carbon or oxygen. The abnormalitystring for this fragment must contain two terms, one to identify the elements that are not represented by E4 and one to require E4 to be deuterated. The latter term employs the indefinite value "n" as the number of deuteriums to which E4 is attached, since this number can vary for valid responses depending on the element to which E4 corresponds in each case.

The structurespecification for this fragment is:

1'E480. (Dl=n. E4= NOT C,O.)
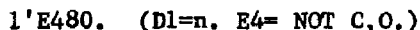
Example 4

[15]N-labeled exocyclic imines are to be retrieved through an A/A search for the fragment
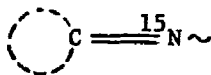
No restrictions are placed on the adjacent atoms or connecting bonds in the ring containing the carbon atom. Therefore in the atomstring for carbon the element symbol C is preceded by as asterisk (*) to require the atom to be a ring member; the unrestricted attachment to the rest of the ring is represented by a type 8 "don't care" bond, which by definition represents any number of any type(s) of cyclic or acyclic bonds. An abnormalitystring is included to require the nitrogen to have a mass number of 15. The structurespecification for this fragment is written

1N22-80. 2*C80. (M1=15.)

Example 5

Consider an A/A search to retrieve derivatives of arsonic acid and thionoarsonic acid represented by the following structure:
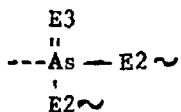
```
        E3
        ‖
---As — E2 ∿          E3 is an atom of O or S
        │
       E2∿            E2 is an atom of N, Se or Te
```

Since the fragment employs the two pseudoelement symbols E2 and E3, the abnormalitystring must include two terms to define these symbols. The structurespecification for this fragment can be written

      1As22-13-14-50. 2E3. 3E280. 4E280. (E3= O,S. E2= N, Se, Te.)

3.5 Part II: Encoded Query

In Part II of the coding form (illustrated below), all of the retrieval requirements that have been specified in Parts III, IV and V are encoded for input.

Illustration: Query Coding Form, Part II

II.
```
ENCODED QUERY:


Query name:
Keydefinitions:      (Section III.)
KEYS =



FORMULA



DEFINE STRUCTURE    (Section V.)
STRUCTURE =




END
```

The encoded query can be viewed as consisting of:

Query name

Query body: Keydefinition(s)

KEYS Logical Statement(s)

FORMULA Statement(s) (Optional)

DEFINE STRUCTURE Statement(s) (Optional)

STRUCTURE Logical Statement(s) (Optional)

END Statement

The _query name_ is an essential part of every query used in output to identify the query to which responses and comments from the system refer.

Each _keydefinition_ identifies to the system one key that the querist may wish to use in the query. At least one keydefinition must appear in every query; usually a number of keys are defined. The fact that a key is defined in a query does not necessarily mean that it _will_ be used, but only that it is available for use. This feature simplifies modification of a query online for browsing or other purposes by enabling the querist to define all potentially useful keys when the query is first input.

The _KEYS logical statement_ identifies the combination of keys required to tag each response. At least one such statement must appear in each query, and every key referred to in the KEYS logical statement must have been previously defined in a keydefinition. If a query contains more than one KEYS logical statement, only the last one is used.

In the _FORMULA statement_ (optional) the restrictions on the molecular formula that are stated in Part IV of the coding form are encoded. If more than one FORMULA statement appears in a query, the last one is used.

The _DEFINE STRUCTURE statement_ (optional) contains the structurespecification(s) for the fragments to be located by atom-by-atom search. A number of fragments can be defined in a single DEFINE STRUCTURE statement; if more than one such statement is used, the effect is cumulative, i.e., each new fragment is simply added to the internal list of fragments defined for use in that query. As with keydefinitions, every fragment defined in a query need not be used.

56

The STRUCTURE logical statement (optional) identifies the combination of fragments that potential responses are to be tested for by A/A search. The structurespecification for each of the fragments must have been previously given in a DEFINE STRUCTURE statement. If more than one STRUCTURE logical statement appears in a query, the last one is used.

The END statement marks the physical end of the input for a single query and is therefore essential.

The following sections describe the construction of each of the above parts in a detailed step-by-step fashion. For examples of completely encoded queries as they are input, see Section 4.

3.5.1 Query Name

Every query is assigned a name that identifies the query in error messages and retrieval printouts. It is the first item of the query that is presented to the preprocessor, and is therefore simply typed as the first character string of the query. The query name consists entirely of letters and digits, must begin with a letter and can be up to five characters in length. Special characters (periods, commas, spaces, etc.) are not permitted, nor are the words KEYS or END, as these have other meanings in the query language.

Examples of valid query names are:

```
H2O
LISA
AMIDE
CCL4
Q1234
J
```

Examples of invalid names are:

PROTEIN	(too many characters)
4SUB	(begins with a number)
COST$	(contains a special character)
P NIT	(contains an imbedded space)

3.5.2 Query Body

The query body consists of the encoded requirements for retrieval, and may contain up to five different types of statements. Two of these types (the keydefinitions, and the KEYS logical statement) must appear in every query; the remaining three types (FORMULA, DEFINE STRUCTURE and STRUCTURE logical statements) are optional.

Rules for construction of these statements, explained in the following sections, all follow the same pattern. In each case, a prototype of the statement is given. One makes appropriate substitutions in the prototype to form the statement used in a given query. The conventions followed in the prototypes are explained in Appendix B.

3.5.2.1 Keydefinitions

The keydefinitions identify and assign keydesignations to the keys that the querist may wish to use in the query. The general form of each keydefinition is

$$\text{keydesignation} = \text{code} \quad /$$

where the keydesignation and CIDS code are listed in Part III of the coding form. The combination(s) of keys required to tag retrieved compounds will be expressed in terms of these keydesignations in a subsequent KEYS logical statement.

As discussed in Section 2, overall search efficiency is affected by the ordering of the keydefinitions in the input. It has been suggested that, wherever possible, the optimum ordering of the keys be implied by the keydesignations. Therefore, if the keydesignations possess some obvious sequence (e.g., K1, K2, K3, ...), the corresponding keydefinitions should appear in this order in the input string.

No more than 36 keydefinitions may appear in each query. All of the keys defined need not necessarily be referred to in a KEYS logical statement. For examples of the use of keydefinitions, see Section 3.5.2.2 or Section 4.

58

3.5.2.2 KEYS Logical Statement

The KEYS Logical Statement is used to specify the logical combination of the keys demanded under "Required key assignment" in Part III of the coding form. It is written in terms of multipliers (1,2,...), the connectors AND OR and NOT, and the keydesignations that have been assigned to keys in preceding keydefinitions. Simplifying the KEYS logical statement by using parentheses to require the proper associations is permitted.

The general form of the KEYS logical statement (before inserting parentheses) is:

$$KEYS = minterm_1 \ OR \ minterm_2 \ ... \ OR \ minterm_n \ \$$$

where $n \geq 1$ (i.e., there must be one or more "minterms" in a KEYS logical statement) and $minterm_i$ is of the form

$$\left[NOT\right]\left[q_1\right] name_1 \ \left[AND\right]\left[NOT\right]\left[q_2\right] name_2 \ ... \ \left[AND\right]\left[NOT\right]\left[q_m\right] name_m \quad 1 \leq m \leq 36$$

where $name_i$ is a keydesignation that has been assigned in a preceding keydefinition.

        q_i is an integer indicating the desired multiple of $name_i$. There must be <u>no space</u> between q_i and the corresponding $name_i$.

<u>At least one $name_i$ in each minterm must be non-negated</u>; otherwise, the minterm is invalid.

If no operator is included between two names, AND is assumed. If no multiplier precedes a name, 1 is assumed.

If more than one logical statement for the KEYS appears in a query, the last statement input is used. Every keydesignation that appears in the last KEYS logical statement must have been previously assigned to a key in a keydefinition. Keys that appear in keydefinitions, but whose keydesignations do not appear in the last KEYS logical statement, have no effect on the retrieval.

The interpretation of the operators AND, OR and NOT is such that NOT is strongest, AND is next strongest, and OR is weakest. This assumed relative binding strength prohibits ambiguity, even in the total absence of

59

parentheses. However, users are cautioned about making errors resulting from using the wrong grouping assumptions. For example,

NAME1 OR 3NAME2 AND NOT NAME3 AND NAME4

could be interpreted in a number of different ways depending on where parentheses are _assumed_. Following the rule above, grouping occurs during preprocessing as if parentheses are included as follows:

(NAME1) OR (3NAME2 AND (NOT NAME3) AND NAME4)

Thus it is said that NOT has tightest hold (greatest binding strength) on its operand NAME3, AND has next tighest hold, and OR has weakest. The preprocessor does _not_, for example, interpret the string as:

(NAME1 OR 3NAME2) AND (NOT(NAME3 AND NAME4))

If one wishes to write a logical expression with the latter meaning without using parentheses, he must write:

NAME1 AND NOT NAME3 AND NOT NAME4 OR 3NAME2 AND NOT NAME3 AND NOT NAME4

To illustrate the combined use of keydefinitions and the KEYS logical statement, consider the following retrieval requirement:

Retrieve all primary and secondary hydroxyalkanemonocarboxylic acids. The pertinent keys are

CIDS Code	Structure
A-C=O	
FG80	---C-O
FG81	—C-O
FG94	O ---C-O

The requirement can be stated:

        one or more FG80 and exactly one FG94 and exactly one A-C=O

OR   one or more FG81 and exactly one FG94 and exactly one A-C=O

The requirement in the form of query statements is:

    K1 = FG81/   K2 = FG80/

    K3 = FG94/   K4 = A-C=O/

(continued)

KEYS = 1K1 AND 1K3 NOT 2K3 AND K4 OR

1K2 AND 1K3 NOT 2K3 AND K4 $

Note that <u>exactly</u> one FG94 (keydesignation K3) is retrieved by demanding

...1K3 NOT 2K3...

This technique need not be employed for the A-C=0 key (keydesignation K4), since this key can not be assigned more than once to each compound.

Using parentheses, we can simplify the logical statement as follows:

KEYS = (K1 OR K2) K3 NOT 2K3 K4 $

since the digit "1" and the connector AND are assumed.

3.5.2.3 <u>FORMULA Statement</u> (Optional)

This statement is used to stipulate the range in which the molecular formulas of all retrieved compounds must lie. The exact requirements to be encoded in the statement are specified in Part IV. of the coding form.

The construction of the formula statement is as follows:

$$
\left[\text{FORMULA} \left\{ \begin{array}{l} \text{HILL}/ \\ \text{ADDEND}/ \end{array} \right\} \text{formula/} \quad \text{formula/} \quad ... \quad \text{formula/} \quad \$ \right]
$$

where each <u>formula</u> has the general form

$$
\left[\text{RESTRICTED} \right] \left\{ \begin{array}{l} \text{el} \\ \text{el(count)} \\ \text{el(lb,)} \\ \text{el(,ub)} \\ \text{el(lb,ub)} \end{array} \right\} ... \left\{ \begin{array}{l} \text{el} \\ \text{el(count)} \\ \text{el(lb,)} \\ \text{el(,ub)} \\ \text{el(lb,ub)} \end{array} \right\} \left[\text{CONSTRAINTS el=a*el } \pm b \; ... \; \text{el=a*el } \pm b \right]
$$

in which

<u>el</u> is an element symbol or the general halogen symbol X
<u>count</u> is an exact count for <u>el</u>, equal to 0, 1, 2...
<u>lb</u> is a lower bound for <u>el</u>, equal to 0, 1, 2...
<u>ub</u> is an upper bound for <u>el</u>
<u>a</u> is an integer from 1 to 63. When a = 1, the 1 and the * may be omitted.
<u>b</u> is an integer from 0 to 31. When b equals zero, the term $\pm$ b may be omitted.

61

The formula statement (if one is present) therefore consists of

    (a)   the word FORMULA, followed by

    (b)   HILL/ if the first _formula_ applies to the Hill molform, or ADDEND/ if the first _formula_ applies to the addend molform (the second and subsequent _formulas_ are assumed to apply to addend molforms); followed by

    (c)   One or more terms of the form _formula/_, where each _formula_ consists of

        (1)   the word RESTRICTED if the restricted feature (Sec. 3.3) is used; followed by

        (2)   one or more terms, each of which consists of either an element symbol alone (to require the element to be present without restricting its count), or an element symbol followed by the exact count or the lower and/or upper bound for the element in parentheses; followed by

        (3)   the word CONSTRAINTS and one or more equations of the form given above, if the constraints feature (Sec. 3.3) is used.

    (d)   A dollar sign ($) terminating the formula statement follows the slash that comes after the last _formula_.

In listing the elements in each _formula_, the _el_'s must be in alphabetical order with the exception that C, H, N, O, and X may appear anywhere.

Example:

If the specifications in Part IV. of the coding form are as follows:

IV. MOLECULAR FORMULA STATEMENT						
Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
ADDEND	yes	C H O	 2	15 26 	20 38 	
/////	yes	C H N O	 2 2	6	8	H=2*C+2

62

the resulting FORMULA statement is

> FORMULA ADDEND / RESTRICTED C(15,20) H(26,38) O(2)/
>
> RESTRICTED C(6,8) H N(2) O(2) CONSTRAINTS H = 2*C + 2 /$

The word ADDEND indicates that, in addition to the second and all subsequent sets, the first set of restrictions also applies to an addend molecular formula. Since two sets of restrictions are provided, records of retrieved compounds must contain at least two addend molecular formulas. One of these formulas must contain:

a. from 15 to 20 carbons
b. from 26 to 38 hydrogens
c. precisely 2 oxygens
d. nothing else (because of the word RESTRICTED)

The other formula must contain:

a. from 6 to 8 carbons
b. any number of hydrogens
c. precisely 2 nitrogens
d. precisely 2 oxygens
e. a number of hydrogens equal to twice the number of carbons plus 2
f. nothing else (because of the word RESTRICTED)

3.5.2.4 DEFINE STRUCTURE Statement (Optional)

The purpose of the DEFINE STRUCTURE statement is to supply the search programs with the structurespecification* for each structure to be sought by an A/A search, and to assign to each an arbitrary name† by which each can be represented internally and externally in error messages. The general form of the DEFINE STRUCTURE statement is

> DEFINE STRUCTURE $structurename_1$ = /$structurespecification_1$/
>
> $structurename_2$ = /$structurespecification_2$/...
>
> $structurename_n$ = /$structurespecification_n$/ $

where $\underline{structurename_i}$ and $\underline{structurespecification_i}$ are as provided in Part V of the coding form.

---

* A structurespecification is an encoded linear representation of a chemical structure or structural fragment. It consists of a connection table with or without an abnormalitystring depending on the structure. Rules for writing structurespecifications are given in Sec. 3.4.4.

† Rules for writing these names are contained in Section 3.4.3.

Structurename$_i$ may appear in a following STRUCTURE logical statement.
Slashes must be used for delimiters as shown above. No more than 36 struc-
tures can be defined in all of the DEFINE STRUCTURE statements of a query.
If more than one DEFINE STRUCTURE statement appear in a query, the effect
is cumulative as with the DEFINE KEYS statement. Two structures in the same
query should not be assigned the same structurename, although a structure
can have the same name as a key.

For examples of the use of the DEFINE STRUCTURE statement, see Section 4.

3.5.2.5 STRUCTURE Logical Statement (Optional)

The STRUCTURE logical statement specifies the particular structure or
combination of structures to be tested for by the A/A search, as described
under "Number of Occurrences" in Part V of the coding form. The statement
is written in terms of multipliers, the connectors AND, OR and NOT and the
structure names that have been assigned in previous DEFINE STRUCTURE state-
ments. The general form of the STRUCTURE logical statement is

$$\text{STRUCTURE} = \text{minterm}_1 \text{ OR } \text{minterm}_2 \ldots \text{ OR } \text{minterm}_n \; \$$$

where $n \geq 1$ and each minterm has the form

$$\left[\text{NOT}\right] \left[q_1 \text{ name}_1\right] \quad \left[\text{AND}\right] \left[\text{NOT}\right] \left[q_2\right]\text{name}_2 \ldots \quad \left[\text{AND}\right] \left[\text{NOT}\right] \left[q_n\right]\text{name}_n$$

where name$_i$ is a structurename assigned in a previous DEFINE STRUCTURE
statement, and q_i is an integer giving the required multiple of the structure
name$_i$. There must be no space between name$_i$ and its multiplier q_i.

The STRUCTURE logical statement is generally similar to the KEYS logical
statement (Sec. 3.5.2.2), except that (a) while a KEYS logical statement is
essential to every query, a STRUCTURE logical statement is not; and (b) while
at least one key in each KEYS minterm must not be negated, it is permitted to
negate all of the structures in a STRUCTURE minterm. Otherwise, the rules
for writing STRUCTURE logical statements are the same as those for writing
KEYS logical statements.

If more than one STRUCTURE logical statement appears in a query, the
last one is used. All of the structures referred to in the statement must

64

have been previously defined in a DEFINE STRUCTURE statement. Structures that are defined in **DEFINE STRUCTURE** statements but that are not referred to in the last STRUCTURE logical statement have no effect on the retrieval.

3.5.3 END Statement

The word END must be included as the last word in the query. Its purpose is to mark the physical end of the query and terminate preprocessing.

The END statement must be punched in columns 1-3 of a card in the batch system and typed at the start of a line in the remote console system. This is the only deviation from the "free format" scanning method used in the system.

A dollar sign may never appear in column 1 of a card in the batch system, since this would signal the end of a set of queries.

4. Examples: Questions with Encoded Queries

The remainder of this document consists of some twenty sample questions typical of those that might be addressed to a working CIDS. All of these examples are within the scope of the early stages of the model system, that is, they probe the structure and/or molecular formula of classes of compounds currently admitted to the file. As additional types of compounds are admitted to the file and as the system's capability for searching non-structural chemical features is implemented, the coding rules will, of course, require updating. However, the considerations that must be taken into account for efficient use of the model CIDS, as illustrated in these examples, will still be applicable even while the system is expanding.

The presentation of each example consists of the following:

(1) The original question as posed by the user.

(2) A brief comment on the data supplied in the original question, identifying the structural and/or molecular formula features that characterize true responses, and determining whether the required compounds can best be retrieved by using more than one query.

(3) A description of the search strategy for each query explaining the selection of the particular combination of search components employed.

(4) An illustration of the exact text of the query as it is input. In the real-time model system, input is via teletype. Once the text of a query has been input, commands can be typed in instructing the system to make alterations in the text or to begin processing the query. The operation and use of the various teletype commands is described in Appendix A.

(5) The completed coding form for each query.

The strategies embodied in these queries have been devised by persons having extensive experience with CIDS to illustrate techniques for using the system most efficiently. However, a user having adequate chemical knowledge and a basic understanding of the system can conduct searches which, although perhaps not maximally efficient are nevertheless satisfactory, even if he lacks sufficient experience to recognize all available means for improving search economics.

66

EXAMPLE 1

Question

What compounds are on file that contain C, H, O, P and either Cd or Mg?

Comment

The desire. family of compounds is characterized not by any common structural feature, but only by the set of element types which may appear in the Hill molecular formula. All responses must contain either C, H, O, P and Cd only or C, H, O, P and Mg only. Such limitations can be imposed in CIDS by use of the Molecular Formula Statement with the RESTRICTED option. Since each Formula statement can impose only one set of restrictions on the Hill formula and since each query may contain only one Formula statement, two queries, EG1A and EG1B, are required to best answer this question.

Strategy for queries EG1A and EG1B

The molecular formula of all retrieved compounds is required to contain any number of each of the elements C, H, O, P and Cd (or Mg). Although these requirements can be completely stated with a Formula statement, each query must first demand the presence of one or more keys in order to obtain a list of potential responses on which the restrictions of the Formula statement can be imposed.

Since the required family of compounds contains no common structural features, no structural fragment keys can be employed. A list of potential responses can be obtained by using three of the molecular formula keys. To require the presence of Cd (or Mg) and P regardless of the number of times each occurs, the qualitative molecular formula keys MF CD (or MF MG) and MF P are demanded. CIDS does not provide a qualitative key for oxygen, since the list for such a key would be exceedingly long. However, the qualitative presence of oxygen in compounds which have responded to the other keys in the query can be required by demanding that the key MF O O not occur.

The keydesignations K1, K2 and K3 are assigned to the cadmium (or magnesium), phosphorus and oxygen keys respectively in accordance with the expected order of increasing length of the keylists.

67

The molecular formula keys have demanded the presence of Cd (or Mg), O and
P. To require that both C and H be present and to assure that no additional element
types are present, the molecular formula statement with the RESTRICTED option is
employed. Since only the qualitative presence of these five elements is required,
no counts (exact, or lower or upper bound) are specified.

Encoded Query

The queries are assigned the names EG1A and EG1B. The input for EG1A is

EG1A

K1 = MF CD/   K2 = MF P/   K3 = MF O Ø/

KEYS = K1 K2 NOT K3 $

FORMULA HILL / RESTRICTED C H CD O P / $

END

The input for query EG1B is

EG1B

K1 = MF MG/   K2 = MF P/   K3 = MF O Ø/

KEYS = K1 K2 NOT K3 $

FORMULA HILL / RESTRICTED C H MG O P / $

END

CIDS Query Coding Form

Query name: _EGIA_

I. **QUESTION:**

What compounds are on file that contain C, H, O, P, and either Cd or Mg?

Structural Representation	Molecular Formula Specifications
	Contains C, H, O, P and Cd; no other elements

II. **ENCODED QUERY:**

Query name: _EGIA_

Keydefinitions: (Section III.)

KEYS =   K1 K2 NOT K3 $

FORMULA _HILL / RESTRICTED   C H CD O P / $_

DEFINE STRUCTURE   (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C) Extracyclic (EC) Number of Cyclic Nuclei (NCN)	K1	MF CD	one AND
Cyclic Nuclei: non-H Attts. (DACN) Generic Cyclic Nuclei (GCN)	K2	MF P	one AND
Specific Cyclic Nuclei (SCN) Specific Functional Group (FG) Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR) Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)	K3	MF O ∅	NOT

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H Cd O P				

70

Query name: *EG1B*

I. QUESTION: *See query EG1A*

Structural Representation	Molecular Formula Specifications

II. ENCODED QUERY:

Query name: *EG1B*

Keydefinitions: (Section III.)

KEYS = *K1   K2   NOT   K3 $*

FORMULA *HILL / RESTRICTED   C   H   MG   O   P / $*

DEFINE STRUCTURE   (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	*K1*	*MF MG*	*one*
Acyclic-Cyclic (A-C)			*AND*
Extracyclic (EC)			*one*
Number of Cyclic Nuclei (NCN)	*K2*	*MF P*	*AND*
Cyclic Nuclei: non-H Attmts. (DACN)			*NOT*
Generic Cyclic Nuclei (GCN)	*K3*	*MF O Ø*	
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	*yes*	*C* *H* *MG* *O* *P*				
/////						

72

EXAMPLE 2

Question

Retrieve all unsubstituted monohydric acyclic alkanols that are not primary in character and that contain from 5 to 8 C atoms.

Comment

The required family of alcohols have molecular formulas in the C_5 to C_8 range, and contain a secondary or tertiary hydroxyl as the only functional group.

Strategy

The key A-C=0 is used to require all responses to be totally acyclic.

Secondary and tertiary alcohol groups which are not attached to a ring are tagged with the specific functional group keys FG81 and FG82 respectively; true responses have been assigned one of these keys exactly once.

$$-C-0 \qquad\qquad \underset{|}{\overset{|}{-C}}-0$$

FG81   FG82
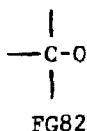
The only heteroatom in the molecular formula of true responses is the single oxygen atom of the hydroxy group. Either the molecular formula statement or the molecular formula key MF 0 1 could be used to specify this information. The list length of MF 0 1 is expected to be quite long, especially in relation to the number of compounds which respond to the structural fragment keys being used. Since the MF 0 1 key is not expected to substantially reduce the number of responses to the keys, and since this query requires a molecular formula statement anyhow, it is more efficient to specify the oxygen count in the formula statement.

A-C=0 is the longest key used in the query and is therefore assigned a higher key designation (larger K number) than either of the keys with which it is intersected. Since FG81 and FG82 are not intersected with each other, it is immaterial which of the two receives the lower keydesignation.

In conjunction with the above, the restrictions specified in the molecular formula statement guarantee that all retrievals are true answers to the original question. All members of the required family of compounds have the general formula $C_nH_{2n+2}O$ (n = 5-8), which can be specified in the formula statement as follows:

73

(a) The RESTRICTED option is employed to assure that only the element types C, H and O are present;

(b) the CONSTRAINTS option is used to require the hydrogen count to be twice the carbon count plus two.

(c) The carbon count is required to fall in the range 5 through 8, while oxgen is required to have an exact count of one. The hydrogen count has already been expressed in terms of the carbon count and need not be stated explicitly.

Encoded Query

The query is assigned the name EG2. The "n and NOT n+1" strategy is used to require responses to have been assigned either FG81 or FG82 exactly one time. The input for query EG2 is

```
EG2
K1 = FG81/    K2 = FG82/    K3 = A-C=∅/
KEYS = K1 NOT 2K1 K3 OR K2 NOT 2K2 K3 $
FORMULA HILL / RESTRICTED C(5,8) H O(1) CONSTRAINTS H = ?*C + 2 / $
END
```

I. **QUESTION:** Retrieve all unsubstituted monohydric acyclic alkanols that are not primary in character and contain from 5 to 8 carbon atoms.

Structural Representation	Molecular Formula Specifications
$m = 3-6$, any carbon configuration $$C - (C_m) - C$$ $$\quad\quad\; | \quad$$ $$\quad\quad OH$$	$$C_m H_{2m+2} O$$ $$m = 5-8$$
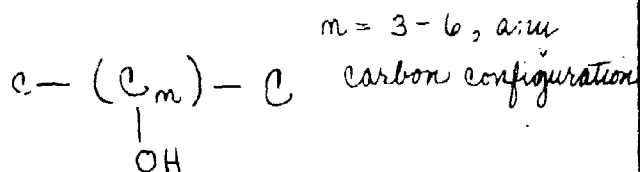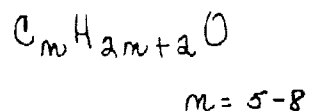
II. **ENCODED QUERY:**

Query name: $EG\,2$

Keydefinitions: (Section III.)

KEYS =
```
K1  NOT  2K1  K3   OR
K2  NOT  2K2  K3   $
```

FORMULA   HILL / RESTRICTED   C(5,8)   H   O(1)
                 CONSTRAINTS   H = 2 * C + 2 / $

DEFINE STRUCTURE   (Section V.)
STRUCTURE =

END

75

III. | KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	KI	FG81	*exactly one*
Acyclic-Cyclic (A-C)			OR
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	K2	FG82	*exactly one*
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)			AND
Specific Cyclic Nuclei (SCN)	K3	A-C=∅	*one*
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. | MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H O	1	5	8	$H = 2 * C + 2$

EXAMPLE 3

Question

Retrieve all compounds that contain both a guanidine residue and a C-formyl group.

Comment

The question demands two characteristic structural fragments: the formyl group (bonded to carbon) and the guanidine residue. The fact that the user specifies a minimum of three N, but does not mention the necessary one or more O has no bearing on whether or not this information will be explicitly stated in the query.

Strategy

In terms of the CIDS keys, the structures of all responses must contain

(1) guanidine in any of the three environments tagged by CIDS keys, namely,
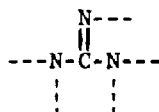
```
        N                        N---                         N ~
        ‖                         ‖                           ‖
     -N-C-N                  ---N-C-N---                  ~ N-C-N ~
                                 ┊   ┊
                                 ┊   ┊

  FG74 or FG74R            FG75 or FG75R                FG76 or FG76R
```
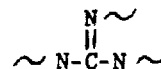
Since each of these fragments may or may not be attached to a ring, a total of six keys are involved. If any one of these keys occurs one or more times, the guanidine requirement is satisfied.

(2) a formyl group attached to C, which is tagged in CIDS by two keys (FG85 and FG85R) depending on whether or not the group is attached to a ring. Occurrence of either of these keys one or more times satisfies the formyl group requirement.

Since the guanidine keys are expected to be shorter than the formyl keys, the lower keydesignations (K1 to K6) are assigned to the guanidines.

It is also known that the molecular formula of all responses will contain C, H, three or more N and one or more O. Since all compounds which respond to both a formyl key and a guanidine key will meet these requirements, restatement of this information in terms of either the MF keys or the MF statement is useless.

77

Encoded Query

The query is assigned the name EG3. The input for this query is

EG3

K1 = FG74/   K2 = FG74R/   K3 = FG75/   K4 = FG75R/

K5 = FG76/   K6 = FG76R/   K7 = FG85/   K8 = FG85R/

KEYS = (K7 OR K8) AND (K1 OR K2 OR K3 OR K4 OR K5 OR K6) $

END

CIDS Query Coding Form

Query name: EG3

I. QUESTION: Retrieve all compounds that contain both a guanidine residue and a C-formyl group.

Structural Representation

N
‖
N − C − N residue regardless of attachments; also

O
‖
C − C − H regardless of attachments
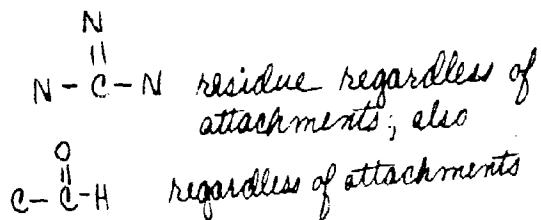
Molecular Formula Specifications

3 or more N atoms

II. ENCODED QUERY:

Query name: EG3

Keydefinitions: (Section III.)

KEYS = (K7 OR K8) AND (K1 OR K2 OR K3 OR K4 OR K5 OR K6) $

FORMULA

DEFINE STRUCTURE (Section V.)
STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key desig-nation	CIDS code	Required assignmt.
Mol. Formule (MF)	K1	F674	one OR
Acyclic-Cyclic (A-C)			
Extracyclic (EC)	K2	FG74R	one OR
Number of Cyclic Nuclei (NCN)			one OR
Cyclic Nuclei: non-H Attmts. (DACN)	K3	FG75	one OR
Generic Cyclic Nuclei (GCN)	K4	FG75R	one OR
Specific Cyclic Nuclei (SCN)	K5	FG76	one OR
Specific Functional Group (FG)	K6	FG76R	one
Nonspec. Diatomics (ND)			AND
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)	K7	FG85	one OR
Metal Cation (CN)			
Inorganic Anion (AN)	K8	FG85R	one
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

EXAMPLE 4

Question

What $C_{10}H_{16}O_2$ compounds are on file that are esters of carboxylic acids and contain a cyclopentene nucleus?

Comment

All responses must have the exact Hill molecular formula $C_{10}H_{16}O_2$. The carboxylate ester residue need not be directly attached to the cyclopentene nucleus.

Strategy

Two structural features are known to characterize true responses: a cyclopentene nucleus (which CIDS tags with the specific cyclic nuclei key SCN32), and exactly one carboxylate ester group. The ester group is tagged by either of two structural fragment keys, FG96 or FG96R, depending on whether or not the group is attached to a ring.
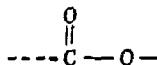
SCN32   FG96 or FG96R

The exact count of C, H and O in the Hill molecular formula could be specified either in the formula statement or with the molecular formula keys MF C 10, MF H 16 and MF O 2, all of which have average or longer keylists. Since the cyclopentene key has a short keylist, the number of responses to all of the structural fragment keys must be relatively small. Therefore, it is faster to test the formula of each response individually with a molecular formula statement than to perform the intersections of the three molecular formula keys.

The cyclopentene key has the shortest keylist and is assigned the lowest keydesignation. Since the two ester keys are "ORed", their keylists are not intersected in this query; therefore, it is immaterial which of the two receives the lower keydesignation, regardless of their relative listlengths.

Since the restriction on the oxygen count in the molecular formula statement eliminates compounds containing more than one ester group, the "n AND NOT n+1" strategy need not be used to require exactly one occurrence of the ester keys.

81

Besides specifying the exact courts of C, H and O in the Hill molecular formula, the molecular formula statement also requires that no additional element types be present by using the RESTRICTED option.

Encoded Query

The query is assigned the name EG4. The input for query EG4 is

```
EG4
K1 = SCN32/    K2 = FG96/    K3 = FG96R/
KEYS = K1 (K2 OR K3) $
FORMULA HILL / RESTRICTED C (10) H(16) O(2) / $
END
```

Query name: EG4

I. QUESTION: What C$_{10}$H$_{16}$O$_2$ compounds are on file that are esters of carboxylic acids and contain a cyclopentene nucleus?

Structural Representation	Molecular Formula Specifications
Must contain $-\overset{O}{\underset{\|\|}{C}}-O-$ and	exactly C$_{10}$H$_{16}$O$_2$

II. ENCODED QUERY:

Query name: EG4

Keydefinitions: (Section III.)

KEYS = K1 AND (K2 OR K3) $

FORMULA HILL / RESTRICTED C(10) H(16) O(2) / $

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

83

III. | KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	SCN 32	one
Acyclic-Cyclic (A-C)			
Extracyclic (EC)			AND
Number of Cyclic Nuclei (NCN)			
Cyclic Nuclei: non-H Attmts. (DACN)	K2	FG 96	one
			OR
Generic Cyclic Nuclei (GCN)	K3	FG 96 R	one
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. | MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C	10			
		H	16			
		O	2			

EXAMPLE 5

Question

Retrieve all pyridinemonocarboxylic acids and esters that contain (a) a total of not more than 12 C atoms, (b) no other heteroatoms, and (c) no additional substitutions on the pyridine nucleus.

Comment

The question defines a structural family (pyridinemonocarboxylic acids and esters) limited to members lying in a particular molecular formula range. Not all of the molecular formula data supplied by the questioner need be explicitly encoded in the query.

Strategy

All true responses must contain:

(1) exactly one occurrence of the pyridine nucleus, which CIDS tags with a specific cyclic nuclei key, SCN44; and,

(2) exactly one $-CO_2$ (carboxyl) group attached to that nucleus, either as the free acid or esterified. The CIDS keys for carboxylic acids and esters attached to ring(s) are, respectively, FG94R and FG96R.
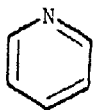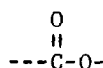
SCN44   FG94R   FG96R

As no heteroatoms other than those in the pyridine ring and the $-CO_2$ group are permitted, the Hill molecular formulas of all retrievals should contain exactly two O's and exactly one N. The molecular formula key lists for MF N 1 and MF O 2 are both very lengthy, however. Since the structural fragment keys alone will severely limit the number of possible responses, it is not worth the computer time required to search on the basis of these MF keys.

Note that the keydesignations K1, K2, ... are assigned in the expected order of increasing length of the key lists.

In the Molecular Formula Statement, the RESTRICTED option can be exercised since all responses must contain only C, H, N, and O. The upper bound of 12 for carbon is specified, along with the exact count of one for N and two for O. Since all responses to the keys will contain the required 6 or more C's, no lower bound for carbon need be specified.

While the structural fragment keys have demanded a $-CO_2$ group attached to a ring, it is not necessarily true that this group is attached to the pyridine ring (since additional hydrocarbon rings can be present). If desired, an atom-by-atom search could be performed to assure such an attachment. Since the carboxyl may be attached to the 2, 3 or 4 postions of the pyridine, the minimum structural fragments which will detect these isomers are
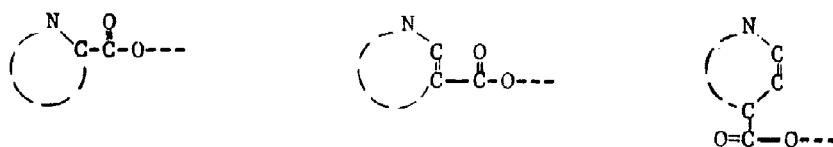
in which all ring bonds are resonant.

For this question, however, the percentage of false responses is expectably low, and their elimination by visual examination of the printout of all responses to the keys and formula statement is a more economical procedure.

Encoded Query

The query is assigned the name EG5. The input for query EG5 is

EG5

K1 = SC1.44/   K2 = FG94R/   K3 = FG96R/

KEYS = K1 (K2 OR K3) $

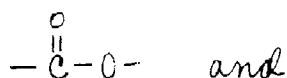FORMULA HILL / RESTRICTED C(,12) H N(1) O(2) /$

END

I.

QUESTION: Retrieve all pyridine monocarboxylic acids and esters that contain (a) not more than a total of 12 C atoms, (b) no other heteroatoms and (c) no additional substitutions on the pyridine nucleus.
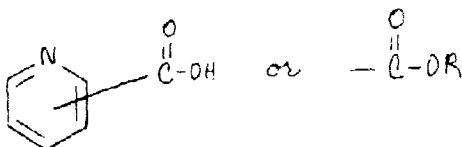
Structural Representation	Molecular Formula Specifications
[pyridine ring structure with substituent] C-OH or -C-OR	Must contain NO_2 and 6-12 C atoms. No other heteroelements.

II.

ENCODED QUERY:

Query name: EG5

Keydefinitions: (Section III.)

KEYS = K1 AND (K2 OR K3) $

FORMULA HILL / RESTRICTED C(,12) H N(1) O(2) / $

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

87

III.

KEYS			
User's Checklist of Key Types	Key desig-nation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C) Extracyclic (EC) Number of Cyclic Nuclei (NCN) Cyclic Nuclei: non-H Attmts. (DACN) Generic Cyclic Nuclei (GCN) Specific Cyclic Nuclei (SCN) Specific Functional Group (FG) Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR) Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)	K1 K2 K3	SCN 44 FG 94 R FG 96 R	one AN one OR mn

IV.

MOLECULAR FORMULA STATEMENT						
Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H N O	 1 2		12	

88

EXAMPLE 6

Each of the three parts a, b and c of example 6 (p. 91) is actually a separate question and will be treated as such in the discussion below.

Question 6a

How many acyclic C-monocarboxylic acids are on file?

Comment 6a

The question is not concerned with outputting a particular family of compounds, but only with obtaining statistics on the presence of that family in the search file. Such statistics are generally useful in probing the quality of sizable files of compounds and in deciding on the desirability of restricting or expanding a contemplated query. In this case, for example, if the number obtained in response to 6a were acceptably small, the structures could be printed out and the answers to 6b and 6c obtained by visual examination.

Strategy

True responses must be totally acyclic, i.e., they must have been assigned the A-C=O key. In addition, they must contain exactly one carboxylic acid group attached to carbon; the specific functional group key for -COOH attached to carbon is FG94.

$$
\begin{array}{c}
O \\
\parallel \\
\text{---C} \text{---} O
\end{array}
$$
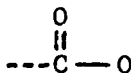
FG94

The design of the real time CIDS is such that when a query is submitted, the system first performs the required combinations of the keys, and prints out the number of items in the accession list, i.e., the number of responses to the keys. Since the family of acyclic monocarboxylic acids is completely described by the above keys, the the total number of compounds in this class which are on file is the number of items in the accession list.

As usual, the lower keydesignation K1 is assigned to the key expected to have the shorter keylist, in this case, FG94.

Encoded Query

The query is assigned the name EG6a. The "n AND NOT n+1" strategy is used to require the carboxylic acid key to have been assigned to each retrieval exactly once. The input for query EG6a i

```
EG6A
K1 = FG94/    K2 = A-C=Ø/
KEYS = K1 NOT 2K1 K2 $
END
```

Query name: $EG6A$

I. QUESTION: (a) How many acyclic c-monocarboxylic acids are on file? (b) Retrieve those that are saturated and contain one or more halogen atom as the only functional group.
(c) Isolate those from (b) that contain only one halogen atom, which must be beta relative to the carboxyl.

Structural Representation	Molecular Formula Specifications
See question	(a) none (b) contains C, H, C and X only (c) $C_m H_{2m-1} O_2 X$

II. ENCODED QUERY:

Query name: $EG6A$

Keydefinitions: (Section III.)

KEYS = ni NC: 2KI h2 $

FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

III.   KEYS

User's Checklist of Key Types				Required assignmt.
Mol. Formula (MF)	FM			
Acyclic-Cyclic (A-C)				
Extracyclic (XC)				
Number of Cyclic Nuclei (NCN)				
Cyclic Nuclei: non-H Attmts. (CNAT)	AA	$A - C - E$		
Generic Cyclic Nuclei (GCN)				
Specific Cyclic Nuclei (SCN)				
Specific Functional Group (FG)				
Nonspec. Diatomics (ND)				
Nonspec. Monatomics (NM)				
Hydrocarb. Radicals (HR)				
Inorganic (IN)				
Metal Cation (MC)				
Inorganic Anion (AN)				
Abnormal Mass (MASS)				
General Metal (MF M)				
Nonstructural (DATA)				
Registry Number (RN)				

IV.   MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

Question 6b

Retrieve those (acyclic C-monocarboxylic acids) that are saturated and contain one or more halogens as the only other functional group.

Comment 6b

The search for halogenated alkanoic acids can be conducted on the basis of many of the structural and molecular formula features which characterize this family of compounds. The object is to determine the minimum set of these features which will successfully retrieve true answers.

Strategy

Structural   features which are common to all members of the required family of compounds are

(a) exactly one carboxylic acid residue not attached to a ring; CIDS tags this fragment with the specific functional group key FG94. $\left(---\overset{\overset{O}{\parallel}}{C}-O \right)$.

(b) the complete absence of rings. Totally acyclic compounds are tagged in CIDS with the key A-C=0.

(c) complete saturation of all carbon-carbon bonds. The key EC1=0 tags structures having zero carbon-carbon double bonds, while the key EC2=0 tags structures having zero carbon-carbon triple bonds.

A single carboxyl is the only oxygen containing functional group permitted; therefore responses must have been assigned the molecular formula key MF 0 2, which has a very long keylist. Combination of the four structural fragment keys as described in (a), (b) and (c) above is expected to obtain a sizable number of responses. Since this number can be substantially reduced by demanding only those compounds containing exactly two oxygens, the key MF 0 2 is used in spite of its long list length. Use of MF 0 2 has an added advantage in that it guarantees that responses have been assigned the carboxyl key exactly once, without having to use the "n and not n+1" strategy.

The only substitutions permitted on this family of acids are one or more halogen atoms. Various techniques are available for stipulating this information.

First, one could use the six specific functional group keys (FG112 to FG117) which tag all mono-, di-, and trihalogenated carbon atoms having no other substitutions. Due to the length of the keylists and the complicated KEYS logical expression that would result, use of these keys is not an efficient method for demanding the required halogenation. A second technique employs the four qualitative molecular formula keys for halogen MF F, MF Cl, MF Br and MF I. Although fewer MF keys than FG keys are required, the MF keys have even longer keylists than the FG keys involved; therefore, use of the qualitative halogen MF keys is also an uneconomical approach. The third technique demands the qualitative presence of the general halogen symbol X in the molecular formula statement. By using the RESTRICTED option to insure that C, H, O and X are the only element types present, and having already required the two oxygen atoms to function in the carboxyl group, it is assured that the halogen(s) present function in the required capacity, i.e., as simple substitutions on carbon atoms. Note that the exact count of two for O, having been stated in the keys, is not repeated in the molecular formula statement.

Encoded Query

The query is assigned the name EG6b. The input for this query is

EG6B

K1 = FG94/   K2 = EC2=∅/   K3 = EC1=∅/

K4 = A-C=∅/   K5 = MF O 2/

KEYS = K1 K2 K3 K4 K5 $

FORMULA HILL / RESTRICTED C H O X / $

END

Query Coding Form

I. QUESTION: *See question on query EG6A*

Structural Representation	Molecular Formula Specifications

II. ENCODED QUERY:

Query name: *EG6B*

Keydefinitions: (Section III.)

KEYS = K1   K2   K3   K4   K5 $

FORMULA *HILL / RESTRICTED*   C H O X / $

DEFINE STRUCTURE   (Section V.)

STRUCTURE =

END

95

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FC-94	one AND
Acyclic-Cyclic (A-C)	K2	≤ C2-1	one AND
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	K3	≤ C1=1	one AND
Cyclic Nuclei: non-H Attmts. (DACN)	K4	A C 1	one AND
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)	K5	MF C 2	one C...
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA·STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H O X				

96

Question 6c

Isolate those (saturated acyclic C-monocarboxylic acids containing one or more halogens as the only other functional group) that contain only one halogen, which is beta relative to the carboxyl.

Comment 6c

Responses to 6c are a special group of the halogenated alkanoic acids which were retrieved in 6b, namely, those acids having exactly one atom of halogen, that halogen being attached to a beta carbon.

Strategy

The structural characteristics of true responses are

(a) the complete absence of rings; responses must therefore have been assigned the key A-C=O.

(b) exactly one carboxylic acid group. The specific functional group key for -COOH not attached to a ring is FG94.

(c) exactly one primary, secondary or tertiary monohalogenated carbon atom, not attached to a ring, which CIDS tags with the specific functional group keys FG112, FG113 and FG114 respectively.

$$---C-X \qquad -\overset{\mid}{C}-X \qquad -\overset{\mid}{\underset{\mid}{C}}-X$$

$$FG112 \qquad\qquad FG113 \qquad\qquad FG114$$

The molecular formula of each response must contain exactly one atom of halogen and two atoms of oxygen. These requirements could be stated by demanding retrievals to have been assigned the key MF O 2 plus one or more of the keys MF F 1, MF CL 1, MF BR 1 or MF I 1. All of these keys have very long list lengths. Since the functional group keys tagging the halogenated carbon (FG112-FG114) have relatively short keylists, it is considered more economical to state the counts of oxygen and halogen in the formula statement than to use these molecular formula keys.

FG112, FG113 and FG114 are the shortest keys used in the query and are therefore assigned the lower keydesignations. Since these three are not intersected with each other in this query, their ordering relative to each other is immaterial.

The limitations imposed on the Hill formula by the molecular formula statement also affect the structural possibilities of responses. Using the RESTRICTED option to assure that the exact set of element types C, H, O and X are present eliminates all compounds having functional groups containing heteroatoms other than O and X. Demanding exactly one X and exactly two O limits the O- and X- containing functional groups to exactly one -COOH and exactly one of the three groups tagged by the keys FG112, FG113 or FG114. Therefore it is not necessary to employ the "n AND NOT n+1" strategy in the keys logical statement to retrieve exactly one of any of these keys. Finally, the CONSTRAINTS option, by demanding the relationship $H = 2*C - 1$ between the hydrogen count H and the carbon count C, insures that all responses are saturated.

If the query were terminated here, the output would consist of all monohalogenated alkanoic acids on file. The original question, however, demands only those acids in which the halogen is attached to the beta carbon. Since the percent of responses to the keys and formula statement which are <u>not</u> true answers to the original question is expected to be substantial, an A/A search for the fragment O-C-C-C-X is conducted to identify all of the beta isomers. Note that since the only oxygen atoms present are those in the carboxyl group, only one of these O atoms need be explicitly searched for.

Encoded Query

The query is assigned the name EG6c. The input for query EG6c is

```
EC6C
K1 = FG112/   K2 = FG113/   K3 = FG114/
K4 = FG94/   K5 = A-C=0/.
KEYS = (K1 OR K2 OR K3) AND K4 K5 $
FORMULA /RESTRICTED C H O(2) X(1) CONSTRAINTS H = 2*C - 1 /$
DEFINE STRUCTURE S1 = /1X12.  2C13-80.  3C14-80.  4C15-80.50./$
STRUCTURE = S1 $
END
```

Query name: $E6$

QUESTION: See EG6A

Structural Representation	Molecular Formula Specifications

ENCODED QUERY:

Query name: $EG6C$

Keydefinitions: (Section III.)

KEYS = (K1 OR K2 OR K3) AND K4 K5 $

FORMULA HILL / RESTRICTED C H O(2) X(1) CONSTRAINTS
$$H = 2 * C - 1 / \$$$

DEFINE STRUCTURE (Section V.)

STRUCTURE = S1 $

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C)	K1	FG-112	one OR
Extracyclic (EC) Number of Cyclic Nuclei (NCN)	K2	FG 113	one OR
Cyclic Nuclei: non-H Attmts. (DACN) Generic Cyclic Nuclei (GCN)	K3	FG 114	one AND
Specific Cyclic Nuclei (SCN) Specific Functional Group (FG)	K4	FG 94	one AND
Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR) Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)	K5	A-C = O	one

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H O X	2 1			$H = 2 * C - 1$

100

V. ATOM-BY-ATOM SEARCH

Structure(s):

$$O - \overset{}{C} - \overset{}{C} - \overset{}{C} - X$$
$$\text{5} \quad \text{4} \quad \text{3} \quad \text{2} \quad \text{1}$$
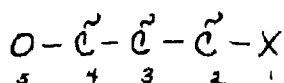
all bonds acyclic

Structure name	Structure specification	No. of Occur.
S1	1 X 1 2. 2 C 1 3 - 8 Ø. 3 C 1 4 - 8 Ø. 4 C 1 5 - 8 Ø. 5 O.	onc.

EXAMPLE 7

Question

Retrieve all fluocarbons within the homologous benzenoid series C_7H_8—$C_{10}H_{14}$.

Comment

True responses are derivatives of hexafluorobenzene in the C_7- C_{10} formula range in which one or more of the fluorine atoms is replaced by a perfluoroalkyl radical.

Strategy

Since each member of this series of compounds contains a benzene nucleus as the only ring system, retrievals are required to have been assigned both of the structural fragment keys A-C=1 (which tags monocyclic compounds) and SCN48 (which tags benzene nuclei).

Each substituting perfluoroalkyl radical terminates in a -CF_3 group. If the substitution is the trifluoromethyl radical, the -CF_3 is attached directly to the benzene ring; in larger radicals, attachment is to an acyclic carbon atom. CIDS does not tag the -CF_3 group specifically;

instead, the fragment
```
     X
     |
---C-X
     |
     X
```
in which the X's may be any three halogen atoms is tagged with either of the specific functional group keys FG117 or FG117R depending on whether or not the group is attached to a ring.

SCN48
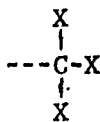
```
     X
     |
---C-X
     |
     X
```
FG117 or FG117R

Since none of the structural fragment keys requires fluorine in the Hill molecular formula, the qualitative molecular formula key MF F is demanded. Use of this key is considered preferable to using the formula statement for demanding fluorine because the relatively short list length of MF F is expected to eliminate a substantial number of the compounds which respond to the other keys in the query.

102

In the molecular formula statement, the RESTRICTED option is exercised to require C and F to be the only element types in the Hill formula, while the constraint F = 2*C — 6 necessitates the required relationship between the counts of carbon and fluorine. Although the user has noted that from seven to ten carbon atoms are required, only the upper bound of ten need be explicitly stated since all responses to the structural fragment keys contain at least seven C's. Since the fluorine count has been expressed in terms of the carbon count, no lower or upper bound for this element is specified.

Encoded Query

    The query is assigned the name EG7. The input for query EG7 is
    EG7
    K1 = FG117/ K2 = FG117R/ K3 = MF F/
    K4 = A-C=1/ K5 = SCN48/
    KEYS = (K1 OR K2) K3 K4 K5 $
    FORMULA HILL / RESTRICTED C(,10) F CONSTRAINTS F = 2*C - 6 /$
    END

I. QUESTION: *Retrieve all fluocarbons within homologous benzenoid series* $C_7F_8 - C_{10}F_{14}$.

Query name: EG 7

Structural Representation	Molecular Formula Specifications
See question.	$C_m F_{2m-6}$

II. ENCODED QUERY:

Query name: EG 7

Keydefinitions: (Section III.)

KEYS = (K1 OR K2) AND K3 K4 K5 $

FORMULA HILL / RESTRICTED C (,10) F CONSTRAINTS

$$F = 2 * C - 6 / \$$$

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FG 117	one ⎫
Acyclic-Cyclic (A-C)			OR ⎬
Extracyclic (EC)			one ⎭
Number of Cyclic Nuclei (NCN)	K2	FG 117 R	AND
Cyclic Nuclei: non-H Attmts. (DACN)	K3	MF F	one AND
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)	K4	A-C = 1	one AND
Specific Functional Group (FG)	K5	SCN 48	one
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C F			10	$F = 2 * C - 6$

EXAMPLE 8

Question

Retrieve any and all prega-4-enols and pregn—1,4-dienols with or without additional functional groups.

Comment

All members of the two required families of steroids contain a minimum of 21 carbons and 1 oxygen, although no molecular formula specifications have been explicitly stated by the questioner. Note that additional substitutions on the steroid nucleus are permitted.

Strategy

The required structural features in terms of the CIDS keys are:

(1) one or more occurrences of either the $\triangle^4$-steroid nucleus or the $\triangle^{1,4}$-steroid nucleus; CIDS employs specific cyclic nuclei keys for each of these nuclei, SCN130 and SCN133 respectively.

(2) an -OH group on any of the 21 carbon atoms in the basic structural fragment. The pertinent CIDS keys are FG80 (if the -OH is attached to C^{21}), FG80R (if attached to C^{18} or C^{19}), FG81R (if attached to C^{20}), or FG83 (if attached directly to the nucleus). Any one of these four keys occurring one or more times will satisfy the -OH requirement.

Since the -OH is permitted to be attached to carbons 18 through 21, no hydrocarbon radical keys can be stipulated.

In the molecular formula statement, a lower bound of 21 carbons is demanded, since this is not necessitated by the keys. A minimum of one O need not be stipulated since all responses to the -OH keys will meet this requirement.

106

SCN130

SCN133

```
---C-O          --C-O              C-O
```
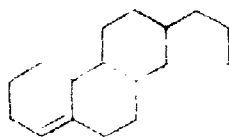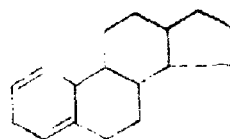
FG80 or FG80R   FG81R   FG83

Encoded Query

The query is assigned the name EG8. The input for this query is

EG8

K1 = SCN130/   K2 = SCN133/   K3 = FG80/

K4 = FG80R/   K5 = FG81R/   K6 = FG83/

KEYS = (K1 OR K2) (K3 OR K4 OR K5 OR K6) $

FORMULA HILL /C(21,)/$

END

Query name: *EG 8*

I. QUESTION: Retrieve any and all pregn-4-enols and pregna-1,4-dienols with or without additional functional groups.

Structural Representation
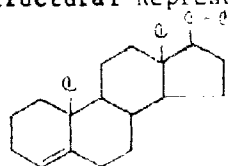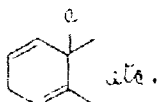
or etc.

one or more —OH groups anywhere; additional functional groups permitted.

Molecular Formula Specification Form

II. ENCODED QUERY:

Query name: *EG 8*

Keydefinitions: (Section III.)

KEYS = (K1 OR K2) (K3 OR K4 OR K5 OR K6) $

FORMULA HILL | C(21,) | $

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

KEYS			
User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C)	K1	SCN 130	one OR
Extracyclic (EC) Number of Cyclic Nuclei (NCN)	K2	SCN 133	one
Cyclic Nuclei: non-H Attmts. (DACN)			AND
Generic Cyclic Nuclei (GCN)	K3	FG 80	one OR
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)	K4	FG 80 R	one OR
Nonspec. Diatomics (ND) Nonspec. Monatomics (NM)	K5	FG 81 R	one OR
Hydrocarb. Radicals (HR) Inorganic (IN)	K6	FG 83	one
Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL		C		21		

EXAMPLE 9

Question

Retrieve all derivatives of o-, m-, and p- cresol in which the only additional functional group is (a) either nitro or amino or acylamino and (b) ortho- to the hydroxyl function.

Comment

While the individual structural features which characterize true responses (e.g., benzene ring, methyl group, amino group, etc.) are relatively common, compounds containing the required combination of these features in the proper juxtaposition constitute a very restricted family.

Strategy

The structural features which must characterize retrievals are

(a) a benzene ring, which CIDS tags with the specific cyclic nuclei key SCN48;

(b) a hydroxyl attached to the ring, which is tagged by the specific functional group key FG83 ;

(c) a methyl group attached to the ring, tagged by the hydrocarbon radical key HR1R;

(d) either an amino group, a nitro group or an amido group attached to the ring. The specific functional group keys for these fragments attached to a ring are FG143R, FG154R and FG35R respectively.

The molecular formula of each true response contains exactly one atom of N and from one to three atoms of O (one in the case of aminocresols, two for amidocresols and three for nitrocresols). Since the structural fragment keys in (a) through (d) above limit the number of potential responses to a relatively small number, and since the molecular formula keys involved (MF N 1, MF O 1, MF O 2 and MF O 3) all have lengthy keylists, it is more efficient to state the counts of N and O in the formula statement than by using molecular formula keys.

110

SCN48	FG83	HR1R	

$$-N$$

FG143R

$$\overset{O}{\overset{\|}{-C-N-}}$$

FG35R

$$\overset{O}{\overset{\|}{-N=O}}$$

FG154R

C-R

The three nitrogen-containing keys are all shorter than any of the cresol keys (i.e., the phenyl, methyl and hydroxy keys), and are therefore assigned the lower keydesignations K1, K2 and K3. Since none of the three nitrogen keys are intersected with each other, their ordering relative to each other is immaterial.

The molecular formula statement proves very useful in eliminating compounds which contain functional groups in addition to those identified in the question. Using the RESTRICTED option to limit the set of element types in the Hill molecular formula to C, H, N and O eliminates compounds having functional groups which contain heteroatoms other than N and O. Demanding exactly one N assures that the only nitrogen containing functional group present is the single amino, amido or nitro group which the user has required. Stating the permitted range of one to three oxygen fails most, although not necessarily all compounds with oxygen-containing functional groups other than those permitted.

To insure that the phenolic -OH is ortho to the nitrogen-containing functional group, an A/A search is conducted for the fragment

Ring resonant (type 4) bond
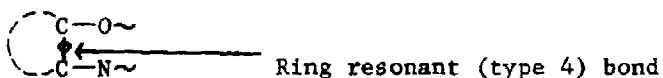
Note that only the smallest fragment which will identify the required isomer is tested for. Since a wiggly bond represents any number of any type bonds or no bond at all, the nitrogen atom will match both the trivalent N in the amino and amido groups and the pentavalent N in the nitro group.

Encoded Query

The query is assigned the name EG9. The input for this query is

```
EG9
K1 = FG154R/   K2 = FG143R/   K3 = FG35R/
K4 = FC83 /    K5 = SCN48/    K6 = HR1R/
KEYS = (K1 OR K2 OR K3) K4 K5 K6 $
FORMULA HILL / RESTRICTED C H N(1) O(1,3) / $
DEFINE STRUCTURE S1 = /1N12-80.  2*C4*3-80.  3*C14-80.  40./ $
STRUCTURE = S1 $
END
```

I.

QUESTION: *Retrieve all derivatives of o-, m-, and p-cresol in which the only additional functional group is (a) either nitro or amino or acylamino and (b) ortho to the hydroxyl function.*

Structural Representation

CH₃
OH
NO₂ or NH₂ or NHCOR
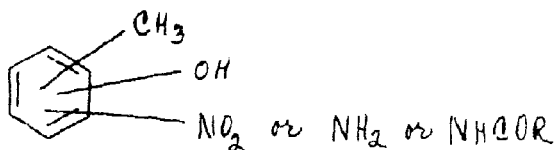
Molecular Formula Specifications

Contains C, H, N and O only

II. **ENCODED QUERY:**

Query name: *EG 9*

Keydefinitions: (Section III.)

KEYS = $(K1\ OR\ K2\ OR\ K3)\ AND\ K4\ K5\ K6\ \$$

FORMULA HILL | RESTRICTED C H N(1) O(1,3) / \$

DEFINE STRUCTURE (Section V.)

STRUCTURE = S1 \$

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FG154 R	mc ⎫
Acyclic-Cyclic (A-C)			or ⎪
Extracyclic (EC)			⎪
Number of Cyclic Nuclei (NCN)	K2	FG148 R	or ⎬
			or ⎪
Cyclic Nuclei: non-H Attmts. (DACN)	K3	FG35 R	or ⎭
Generic Cyclic Nuclei (GCN)			AND
Specific Cyclic Nuclei (SCN)	K4	FG 3	or
Specific Functional Group (FG)			or or
Nonspec. Diatomics (ND)	K5	SCN 8	or
Nonspec. Monatomics (NM)			AND
Hydrocarb. Radicals (HR)			
Inorganic (IN)	K6	JR R	or
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C H N O	1	1	3	
/////						

114

V. ATOM-BY-ATOM SEARCH

Structure(s):

the C-C bond
is resonant (type 4)

Structure name	Structurespecification	No. of Occurrences
S1	1N12-80. 2*C4*3-80. 3*C14-80. 4O.	one

EXAMPLE 10

Question

Retrieve all phosphonothiolic acid esters that are members of the family structured below:

$$
\begin{array}{c}
O \\
\parallel \\
R-P-S-(CH_2)_n-N-R'' \\
\mid \qquad\qquad\;\; \mid \\
O-R' \qquad\;\; R'''
\end{array}
$$

$R = CH_3$ or C_2H_5

R' = any 2-4 carbon alkyl

$R'' = R''' = CH_3$, C_2H_5 or $\underline{n}-C_3H_7$

Comment

The question demands retrieval of a family of compounds which differ from each other in the hydrocarbon radicals attached to the functional groups. The substitutions permitted at each location are stated exactly, and a general molecular formula is provided.

Strategy

The A-C=O key immediately limits responses to totally acyclic compounds. In addition to being acyclic, every response must contain one phosphonothiolic ester group and one tertiary amine group. The specific functional group keys for these two fragments not attached to rings are FG202 and FG145.

The ester group is linked to the amino nitrogen by one of the three radicals ethylene, trimethylene or tetramethylene. The hydrocarbon radical keys for these fragments attached to two heteroatoms are, respectively, HR6EE, HR14EE and HR24EE.

$$
\begin{array}{ccccc}
O & & & & \\
\parallel & & & & \\
---P-S--- & \quad -N- & \quad \sim E\ell-C-C-E\ell\sim & \quad \sim E\ell-C-C-C-E\ell\sim & \quad \sim E\ell-C-C-C-C-E\ell\sim \\
\mid & & & & \\
O--- & & & & \\
\\
FG202 & \quad FG145 & \quad HR6EE & \quad HR14EE & \quad HR24EE
\end{array}
$$

The only additional structural fragments in true responses are the alkyl radicals attached to the ester and amino groups. Every radical permitted to occur in true responses is tagged in CIDS by a hydrocarbon radical key; therefore, a query could be written demanding each combination of structural fragment keys which characterizes a true response. However, such an approach would result in a very long and cumbersome query.

Analysis of the question reveals that each response must contain at least one methyl or one ethyl group attached to a heteroatom. In view of the short keylist of the phosphonothiolate ester group, it is judged sufficient to cite only the methyl and ethyl keys explicitly. The remaining radicals will be limited by restricting the size of the molecular formula, with final identification of true responses by examining the output.

```
        C-El~                    C-C-El~
        HR1E                     HR3E
```

Every response has exactly one atom of N, one P, one S and two O, and thus would have been assigned the molecular formula keys MF N 1, MF P 1, MF S 1 and MF O 2. Since, however, all of these keys have very long keylists, especially relative to the phosphonothiolate key, it is more efficient to state the formula restrictions in the molecular formula statement than with these MF keys.

Keydesignations are assigned in the expected order of increasing length of the keylists. Since the phosphonothiolate key has such a short keylist and since the restrictions in the formula statement further limit the fragments that can occur in responses, it is not necessary to use the "n AND NOT n+1" strategy to demand exact occurrences of any of structural fragment keys.

Besides stating the exact count for N, O, P and S, the molecular formula statement also demands that C and H be the only additional element types present. The permitted range of 7 to 14 carbon atoms is stated, and the relationship H = 2*C + 4 between the hydrogen count H and the carbon count C is required. Since the H count is thereby expressed in terms of the carbon count, no count for hydrogen need be stated explicitly.

Encoded Query

The query is assigned the name EG10. The input for query EG10 is

```
EG10
K1 = FG202/  K2 = FG145/  K3 = A-C=0/  K4 = HR24EE/
K5 = HR14EE/  K6 = HR6EE/  K7 = HR3E/  K8 = HR1E/
KEYS = K1  K2  K3  (K4 OR K5 OR K6)  (K7 OR K8) $
FORMULA HILL / RESTRICTED C(7,16) H N(1) O(2) P(1) S(1)
            CONSTRAINTS H = 2*C + 4 / $
END
```

I.

QUESTION: Retrieve all phosphonothiolic acid esters that are members of the family structured below.

Structural Representation	Molecular Formula Specifications

Structural Representation:

$$R - \overset{\overset{O}{\|}}{\underset{\underset{O-R'}{|}}{P}} - S - (CH_2)_m - N \overset{R''}{\underset{R'''}{<}}$$

$m = 2-4$

$R = CH_3$ or C_2H_5 $R' = $ any 2-4 carbon alkyl

$R'' = R''' = CH_3,\ C_2H_5$ or $m-C_3H_7$

Molecular Formula Specifications:

$C_m\ H_{2m+4}\ N O_2 P S$

II.

ENCODED QUERY:

Query name: EG10

Keydefinitions: (Section III.)

KEYS = K1 K2 K3 (K4 OR K5 OR K6) AND (K7 OR K8) $

FORMULA HILL / RESTRICTED   C(7,16) H N(1) O(2) P(1) S(1)

CONSTRAINTS   H = 2 * C + 4 / $

DEFINE STRUCTURE   (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FG 202	one
Acyclic-Cyclic (A-C)			AND
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	K2	FG 145	one AND
Cyclic Nuclei: non-H Attmts. (DACN)	K3	A-C = 0	one AND
Generic Cyclic Nuclei (GCN)	K4	HR 24 EE	one
Specific Cyclic Nuclei (SCN)			AND
Specific Functional Group (FG)	K5	HR 14 EE	one OR one
Nonspec. Diatomics (ND)	K6	HR 6 EE	OR one
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			AND
Inorganic (IN)			
Metal Cation (CN)	K7	HR 3 E	one
Inorganic Anion (AN)			OR
Abnormal Mass (MASS)	K8	HR 1 E	one
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C		7	16	$H = 2*C + 4$
		H				
		N	1			
		O	2			
		P	1			
		S	1			

EXAMPLE 11

Question

Identify all Ring Index type structures on file, in any state of hydrogenation, that consist of one molecule of either quinoline or isoquinoline in either ortho or ortho-peri-fusion with benzene.

Comment

The user requires all unsubstituted three-ring nuclei resulting from the fusions described, regardless of the number of double bonds.

Strategy

The key DACN = 0 tags each cyclic structure having no direct attachments other than hydrogen. Structures which are assigned this key therefore consist of a single unsubstituted cyclic nucleus.

Having limited retrieved structures to a single nucleus, the generic cyclic nuclei key GCN2 = 6,6,6 is demanded to require that nucleus to consist of three six-membered rings. Two of these rings have the formula C_6 and are tagged with the key GCN3 = C 6; the third ring has the formula C_5^-N, and is tagged with GCN3 = C5 N1.

Depending on whether the benzene and quinoline (or isoquinoline) systems are ortho- or ortho—peri-fused, the complete skeleton of the fused nuclei will have either of the formulas $C_{12}N$ or $C_{13}N$, and will be tagged with the generic cyclic nuclei keys GCN4 = C 12 N 1 or GCN4 = C 13 N 1 accordingly. Demanding retrievals to have been assigned one of these keys eliminates various structures such as

and

Keydesignations are assigned in the expected order of increasing length of the keylists.

Encoded Query

The query is assigned the name EG11. The input for this query is

EG11

K1 = DACN=∅/   K2 = GCN4= C 12 N 1/   K3 = GCN4 = C 13 N 1/

K4 = GCN2=6,6,6/   K5 = GCN3= C 5 N 1/   K6 = GCN3= C 6/

KEYS = K1 K4 K5 2K6 (K2 OR K3) $

END

I. **QUESTION:** *Identify all Ring Index type structures on file, in any state of hydrogenation, that consist of one molecule of either quinoline or isoquinoline in ortho- or ortho-peri-fusion with benzene.*

Query name: $EG11$

Structural Representation

Examples:

Molecular Formula Specifications

Skeleton molecular formula of nucleus =

$C_{12}N$ or $C_{13}N$

II. **ENCODED QUERY:**

Query name: $EG11$

Keydefinitions: (Section III.)

KEYS = $K1 \quad K4 \quad K5 \quad 2K6 \quad (K2 \ OR \ K3) \ \$$
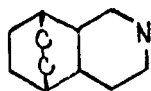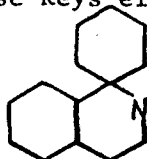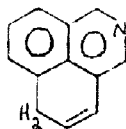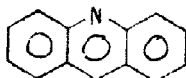
FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

122

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C) Extracyclic (EC)	$K1$	$DACN = 0$	*one* AND *one* ⎫
Number of Cyclic Nuclei (NCN)	$K2$	$GCN4 = C\ 12\ N\ 1$	OR ⎬
Cyclic Nuclei: non-H Attmts. (DACN)	$K3$	$GCN4 = C\ 13\ N\ 1$	*one* ⎭
Generic Cyclic Nuclei (GCN)			AND
Specific Cyclic Nuclei (SCN)	$K4$	$GCN2 = 6,6,6$	*one*
Specific Functional Group (FG)			AND
Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR)	$K5$	$GCN3 = C\ 5\ N\ 1$	*one* AND
Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)	$K6$	$GCN3 = C\ 6$	*two*

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
////						

EXAMPLE 12

Question

What compounds are on file that contain, in any state of hydrogenation, any form of the C_4O-C_6 ring system other than those that have specific individual CIDS codes?

Comment

Retrievals are required to contain any member of the C_4O-C_6 family of cyclic nuclei except those more commonly occurring ones which are tagged with specific CIDS keys.

Strategy

A particular set of cyclic nuclei such as the C_4O-C_6 family can be retrieved by demanding the appropriate set of generic cyclic nuclei (GCN) keys, which tag six types of structural features of ring systems.

All C_4O-C_6 nuclei are assigned the key GCN2=5,6 which tags each cyclic nucleus consisting of one five-membered ring and one six-membered ring. The particular family of nuclei demanded in this query all have the total skeleton molecular formula C_8O, and therefore are tagged with the key GCN4= C8 01. The five-membered C_4O ring in each nucleus is assigned the key GCN3= C4 01.

Like the C_4O ring, the C_6 ring in each nucleus is also tagged with a GCN3 key which could be demanded in this query. The C_6 key also tags such commonly occurring nuclei as benzene, quinoline and naphthalene, and is therefore expected to have a much longer keylist than any of the three GCN keys already demanded. Since the C_6 key is expected to require a considerable amount of time to eliminate only a few false drops, its use in this query is not considered worthwhile.

To eliminate the four C_4O-C_6 nuclei tagged with specific CIDS keys, retrievals are required not to have been assigned the keys SCN83, SCN84, SCN85 and SCN86.
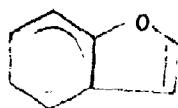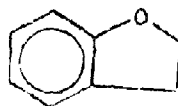
SCN83   SCN84
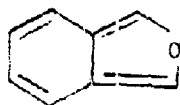
SCN85   SCN86

Key designations are assigned to the keys in expected order of increasing list length.

Since the key GCN3= C8 01 necessitates at least 8 carbons and 1 oxygen in the molecular formula, no molecular formula keys or statement need be employed in this query.

Encoded Query

The query is assigned the name EG12. The input for this query is
EG12
K1 = GCN4= C8 01/   K2 = GCN3= C 4 01/
K3 = GCN2= 5,6/   K4 = SCN83/   K5 = SCN84/
K6 = SCN85/   K7 = SCN86/
KEYS = K1 K2 K3 NOT K4 NOT K5 NOT K6 NOT K7 $
END
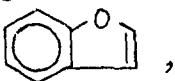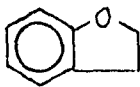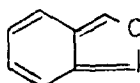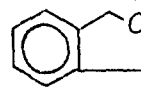
I. | QUESTION: What compounds are on file that contain, in any state of hydrogenation, any form of the C_4O-C_6 ring system other than those that have individual TIDS codes?

Query name: EG12

Structural Representation

Compounds having any C_4O-C_6 ring system except , , or

Molecular Formula Specifications

II. ENCODED QUERY:

Query name: EG12

Keydefinitions: (Section III.)

KEYS = KI K2 K3 NOT K4 NOT K5 NOT K6 NOT K7 $

FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C)	K1	GCN4 = C8 O1	one AND
Extracyclic (EC) Number of Cyclic Nuclei (NCN)	K2	GCN3 = C4 O1	one AND
Cyclic Nuclei: non-H Attmts. (DACN) Generic Cyclic Nuclei (GCN)	K3	GCN2 = 5,6	one AND
Specific Cyclic Nuclei (SCN)	K4	SCN83	NOT AND
Specific Functional Group (FG)	K5	SCN84	NOT AND
Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR)	K6	SCN 85	NOT AND
Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)	K7	SCN86	NOT

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

EXAMPLE 13

Question

Retrieve all salts of N-alkylated isonipicotic acid esters.

Comment

The question requires the distinction of piperidine-4-carboxylic esters from the isomeric -2- and -3- analogs. Only the two substitutions which appear in the structural specifications are permitted on the piperidine ring. Note that the query employs molecular formula data, although none has been provided by the questioner.

Strategy

The required family of compounds contains two characteristic structural fragments:

(1) the piperidine ring, tagged by a specific cyclic nuclei key SCN45, and,

(2) a carboxylate ester group attached to a ring, for which CIDS provides the specific functional group key FG96R.

Since unlimited substitutions are permitted on the R' moiety of the ester, each of the above keys can occur one or more times.

The Hill formula of each true answer reflects not only the substitutions on R', but also the particular organic or inorganic acid addended to the ester. Therefore the only requirements that could be imposed by the MF keys are the qualitative presence of N or O (by demanding not MF N 0 and not MF O O). Since the presence of these elements already results from the more restrictive structural fragment keys employed, it is not necessary to use any molecular formula keys in this query.

SCN45
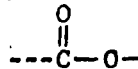
FG96R

The lower keydesignation K1 is assigned to the piperidine, since that ring is expected to have a shorter keylist than the ester group.

In addition to the total (Hill) molecular formula, the records of each true answer must contain an addend molecular formula for the basic (i.e., the piperidine-carboxylate) portion of the total compound and an addend molecular formula for the acidic portion. To eliminate responses to the keys that lack an addend formula, and to help to insure that both the piperidine ring and the ester group are parts of the same structure, a formula statement is employed requiring the presence of an addend molecular formula having the following characteristics:

(1) a minimum of eight carbon atoms (five for the piperidine ring, one for the carboxylate, and one for each of the R and R' substitutions);

(2) a minimum of one N and two O.

To insure that the piperidine ring has only the required substitutions, and that these substitutions are situated at the one and the four positions of the piperidine ring, an A/A search is performed. Note the numbering of the atoms of the fragment. Although all responses to the keys will have a nitrogen atom in a piperidine ring, not all will have a substitution at this atom. Therefore, the N is considered the atom in the fragment least likely to occur in responses to the keys and is assigned number one.

Encoded Query

The query is assigned the name EG13. The input for query EG13 is

```
EG13
K1 = SCN45/   K2 = FG96R/
KEYS = K1 K2 $
FORMULA ADDEND / C(8, )  N(1, )  O(2, ) / $
DEFINE STRUCTURE S1 = /1*N12-1*3-1*10.
2C80.   3*C1*4.   4*C1*5.   5*C16-1*9.
6C27-18.   70.   8010.   9*C1*10.   10*C./$
STRUCTURE = S1 $
END
```

129

I.

QUESTION: *Retrieve all salts of N-alkylated isonipicotic acid esters.*

Query name: *EG-13*

Structural Representation
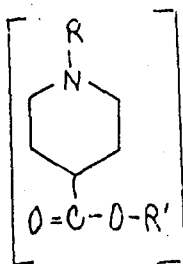
R = any alkyl

R' = any substituted or unsubstituted hydrocarbon radical

· acid

Molecular Formula Specifications

I.

ENCODED QUERY:

Query name: *EG-13*

Keydefinitions: (Section III.)

KEYS = K1 K2 $

FORMULA ADDEND / C(8,) N(1,) O(2,) / $

DEFINE STRUCTURE (Section V.)

STRUCTURE = S1 $

END

III. | KEYS

User's Checklist of Key Types	Key desig-nation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	SCN45	one
Acyclic-Cyclic (A-C)			AND
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	K2	FG96R	one
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. | MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
ADDEND		C		8		
		N		1		
		O		2		

131

V. ATOM-BY-ATOM SEARCH

Structure(s):

Structure name	Structurespecification	No. of Occurrences
51	1 * N 1 2 − 1 * 3 − 1 * 10. 2 C 8 0. 3 * C 1 * 4. 4 * C 1 * 5. 5 * C 1 6 − 1 * 9. 6 C 2 7 − 1 8. 7 0. 8 0 1 0. 9 * C 1 * 10. 10 * C.	one

III. KEYS

User's Checklist of Key Types	Key desig-nation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	SCN45	one
Acyclic-Cyclic (A-C)			AND
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	K2	FG96R	one
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
ADDEND		C	.	8		
		N		1		
		O		2		

V. ATOM-BY-ATOM SEARCH

Structure(s):

Structure name	Structure specification	No. of Occurrences

SI	1*N\2-1*3-1*1∅.
	2C8∅.
	3*C1*4.
	4*C1*5. 5*C16-1*9.
	6C27-18.
	   7∅. 8O1∅. 9*C1*1∅.
	   1∅*C.
		one

EXAMPLE 14

Question

Retrieve isethionic acid and any of its organic or metal salts.

Comment

The user is concerned with an organic acid and compounds in which it functions either in addend or anionic capacity.

Strategy

In terms of the CIDS keys, isethionic acid and its anion consist of the following structural fragments:

(a) either a sulfonic acid group or a sulfonate anion attached to an acyclic carbon, which CIDS tags with the specific functional group key FG223. This same key is also assigned to sulfonic acid esters.

(b) an ethylene group between two heteroatoms, which CIDS tags with the hydrocarbon radical key HR6EE;

(c) a primary alcohol group attached to an acyclic carbon, to which CIDS assigned the specific functional group key FG80.

```
                                              O
                                              ||
   ---C-O          ~E1-C-C-E1~            ---S-O---
                                              ||
                                              O

   FG80              HR6EE                  FG223
```
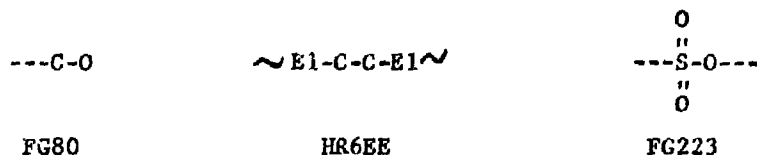
The isethionic acid molecule or the isethionate anion in each true answer is a totally acyclic structure, and responses are therefore required to be assigned the zero cyclic nuclei key NCN=0. This NCN key must be used rather than the A-C=0 key since the NCN keys tag the number of cyclic nuclei in each structure, while the A-C keys are based on the number of rings in a total compound. While true answers to this question must contain the acyclic isethionate structure, they may contain cyclic moieties elsewhere in the compound, thus making A-C≠0.

The utility of including the key NCN=0 in this query might be questioned because its keylist is so much longer than that of the sulfonic acid key. However, in order to fail as many compounds as possible before conducting the A/A search which this query requires, it is preferable to make use of all available keys.

133

Keydesignations are assigned in the expected order of increasing list
length. Since the question places no restrictions on the structure of any
cations or addends which may be present, it is possible for each of these
keys to have been assigned to true answers more than once.

To insure that the structure to which the ethylene, alcohol and isethionic
acid keys are assigned is actually isethionic acid and to eliminate isethionate
esters, an A/A search is conducted for the entire acid,

$$
\overset{\displaystyle O}{\underset{\displaystyle O}{O-C-C-\overset{||}{\underset{||}{S}}-O.}}
$$
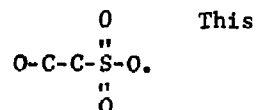
This

fragment retrieves the isethionate anion as well as the free acid, even though
the sulfonic acid group in the fragment is not charged.

Encoded Query

The query is assigned the name EG14. The input for this query is
```
EG14
K1 = FG223/    K2 = FG8Ø/    K3 = HR6EE/    K4 = NCN=Ø/
KEYS = K1 K2 K3 K4 $
DEFINE STRUCTURE   S1 = /1S12-25-26-17 2C13. 3C14. 40. 50. 60. 70./ $
STRUCTURE = S1 $
END
```

Query name: EG 14

I. QUESTION: Retrieve isethionic acid and any of its organic or metal salts.

Structural Representation	Molecular Formula Specifications

Structural Representation:

HO-CH₂-CH₂-S-OH   plus added

compounds plus   HO-CH₂-CH₂-S-O⁻
compounds
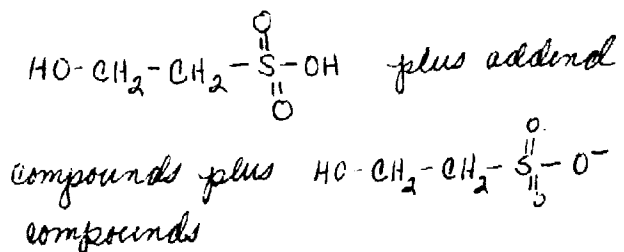
II. ENCODED QUERY:

Query name: EG 14

Keydefinitions: (Section III.)

KEYS = K1 K2 K3 K4 $

FORMULA

DEFINE STRUCTURE   (Section V.)

STRUCTURE = S1 $

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FG223	one AND
Acyclic-Cyclic (A-C)			
Extracyclic (EC)			one
Number of Cyclic Nuclei (NCN)	K2	FG80	AND
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)	K3	HR6EE	one AND
Specific Cyclic Nuclei (SCN)			one
Specific Functional Group (FG)	K4	NCN = 0	
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

136

V. ATOM-BY-ATOM SEARCH

Structure(s):

$$\overset{4}{O}-\overset{3}{C}-\overset{2}{C}-\overset{\overset{\displaystyle O^{5}}{\|}}{\underset{\underset{\displaystyle O_{6}}{\|}}{S}}-O_{7}$$

Structure name	Structurespecification	No. of Occurrences
51	1 S 12−25−26−17. 2 C 13. 3 C 14. 4 0. 5 0. 6 0. 7 0.	one

EXAMPLE 15

Question

Retrieve all alkylpyridinium compounds of the type
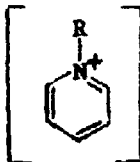
$$\left[\underset{\text{N}^+}{\overset{\text{R}}{\bigcirc}} \right] \cdot \text{ anion (organic or inorganic)}$$

Comment

The user requires all N-alkylpyridinium salts, regardless of the nature of the associated anion. No additional substitutions on the pyridine nucleus are permitted.

Strategy

True responses contain a pyridine nucleus, which CIDS tags with the specific cyclic nuclei key SCN44. This key is assigned regardless of whether or not the nucleus is charged. Responses must also have been assigned the specific functional group key FG146, which tags each positively charged nitrogen atom in a ring.
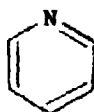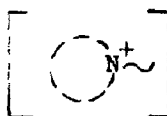
SCN44   FG146

The substitution at the nitrogen atom is always acyclic, and no additional substitutions on the pyridine ring are permitted. Therefore, each retrieval must have been assigned both NCN= 1, which tags structures having exactly one cyclic nucleus, and DACN= 1, which tags structures in which the total number of non-hydrogen attachments to the cyclic nucleus is one. However, since every structure assigned the key DACN= 1 must have exactly one cyclic nucleus, the NCN= 1 key provides no additional information and thus is not demanded in this query.

As evident from the structure, the molecular formula of each true answer must contain at least one N and six C; however, since all responses to the pyridine key contain at least one nitrogen and five carbons and, in virtually all cases, the positive charge on the FG146 nitrogen atom is the result of attachment to carbon, no molecular formula keys or formula statement are required.

The keydesignations K1 through K3 are assigned to the keys in the expected order of increasing list length. Since no restrictions are placed on the anion, all keys are permitted to occur one or more times.

Encoded Query

The query is assigned the name EG15. The input for this query is

```
EG15
K1 = FG146/    K2 = DACN= 1/    K3 = SCN44/
KEYS = K1 K2 K3 $
END
```

Query name: *EG 15*

I. QUESTION: *Retrieve all alkylpyridinium compounds of the type shown below.*

Structural Representation	Molecular Formula Specifications
• anion (organic or inorganic)	
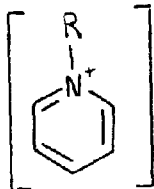
II. ENCODED QUERY:

Query name: *EG 15*

Keydefinitions: (Section III.)

KEYS = K1 K2 K3 ⸴

FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	FG146	one
Acyclic-Cyclic (A-C)			AL.
Extracyclic (EC)			one
Number of Cyclic Nuclei (NCN)	K2	DACN = 1	AND
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)	K3	SCN44	one
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

EXAMPLE 16

Question

Retrieve all compounds of the class structured below in which the hetero-nucleus is either pyrrolidine or piperidine.
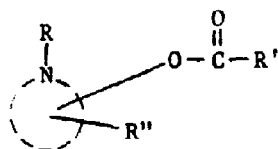
$R=CH_3$ or C_2H_5

R'=any acyclic hydrocarbon radical

R''=cyclopentyl or cyclohexyl

Comment

The user is interested in a typical Markush family of compounds with all features of structural variability clearly stipulated.

Strategy

Retrievals are required to contain the following structural fragments:

(a) either a pyrrolidine or a piperidine nucleus, which are tagged in CIDS with the specific cyclic nucleus keys SCN26 and SCN45, respectively;

(b) either a cyclopentane nucleus or a cyclohexane nucleus; these two specific cyclic nuclei are tagged with the keys SCN31 and SCN49;

(c) a carboxylate ester group attached to a ring, which is tagged with the specific functional group key FG96R;

(d) either a methyl or an ethyl radical attached to the nitrogen of whichever heterocycle (pyrrolidine or piperidine) is present. Since attachment is to a heteroatom that is also a ring atom, the methyl group is assigned both HR1E and HR1R and the ethyl group is assigned both HR3E and HR3R.
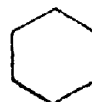
SCN26	SCN45	SCN31	SCN49
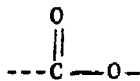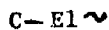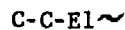
FG96R	HR1E	HR1R	HR3E	HR3R
---C—O-	C—El~	C — R	C-C-El~	C-C-R

Responses are permitted to contain only two cyclic nuclei and are therefore assigned the key NCN = 2. Since these two nuclei have a total of four nonhydrogen attachments, responses are also tagged with the key DACN = 4.

The molecular formula of each true answer has been assigned the molecular formula keys MF N 1 and MF O 2, both of which have very long keylists. Since relatively few compounds are expected to respond to the combinations of structural fragment keys demanded, the molecular formula requirements can be tested for more efficiently with a formula statement than with these two molecular formula keys.

Keydesignations are assigned in expected order of increasing listlength.

In the molecular formula statement, the RESTRICTED option is used to require the Hill formula to contain only the element types C, H, N and O, and the exact counts of one nitrogen and two oxygens are specified. Thus, all compounds containing heteroatoms in addition to the one N and two O are failed, and the R and R' substitutions are automatically hydrocarbon radicals.

Encoded Query

The query is assigned the name EG16. The input for this query is:
EG16
K1 = SCN31/   K2 = SCN49/   K3 = SCN26/   K4 = SCN45/
K5 = FG96R/   K6 = DACN= 4/   K7 = NCN= 2/   K8 = HR1E/
K9 = HR1R/   K10 = HR3E/   K11 = HR3R/
KEYS = (K1 OR K2) (K3 OR K4) K5 K6 K7 (K8 K9 OR K10 K11) $
FORMULA HILL / RESTRICTED C H N(1) O(2) /$
END

I. QUESTION:

Retrieve all compounds of the class structured below in which the nucleus is either pyrrolidine or piperidine.

Structural Representation	Molecular Formula Specifications
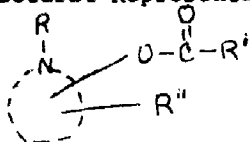
R = CH_3 or C_2H_5
R' = any acyclic hydrocarbon radical
R" = cyclopentyl or cyclohexyl.

II. ENCODED QUERY:

Query name: EG16

Keydefinitions: (Section III.)

KEYS = (K1 OR K2) (K3 OR K4) K5 K6 K7 (K8 K9 OR K10 K11) $

FORMULA HILL / RESTRICTED C 11 N(1) O(2) / $

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

144

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	K1	SCN 31	one ⎫
Acyclic-Cyclic (A-C)			OR ⎬
Extracyclic (EC)	K2	SCN49	one ⎭
Number of Cyclic Nuclei (NCN)			AND
Cyclic Nuclei: non-H Attmts. (DACN)	K3	SCN26	one ⎫ OR
Generic Cyclic Nuclei (GCN)	K4	SCN45	one ⎭
Specific Cyclic Nuclei (SCN)	K5	F696R	AND one AND
Specific Functional Group (FG)	K6	DACN = 4	one AND
Nonspec. Diatomics (ND)	K7	NCN = 2	one
Nonspec. Monatomics (NM)			AND
Hydrocarb. Radicals (HR)			
Inorganic (IN)	K8	HRiE	one
Metal Cation (CN)			AND ⎫
Inorganic Anion (AN)	K9	INiR	one ⎬
Abnormal Mass (MASS)			OR
General Metal (MF M)			
Nonstructural (DATA)	K10	HK3E	one AND ⎫
Registry Number (RN)	K11	HR3R	one ⎭

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yw	C H N O	i 2			
/////						

145

EXAMPLE 17

Question

What N, N-dialkylphenothiazinecarboxamides, with or without additional acyclic ring substituents, are on file?

Comment

Substitutions on the phenothiazine nucleus in addition to the dialkylcarboxamide group must be acyclic, but are otherwise totally unrestricted.

Strategy

Responses are required to have been assigned the specific cyclic nucleus key for the phenothiazine system (SCN116) and the specific functional group key for the disubstituted carboxamide group attached to a ring (FG36R).
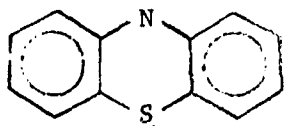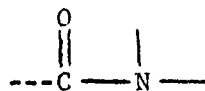
SCN116   FG36R

Since all substitutions must be acyclic, the key NCN = 1 is used to require phenothiazine to be the only cyclic nucleus present in retrieved structures.

Since additional substitutions on the phenothiazine nucleus are permitted, the total number of nonhydrogen attachments to cyclic nuclei in true answers varies, and no DACN key can be specified.

Keydesignations are assigned to the three structural fragment keys in the expected order of increasing list length.

Encoded Query

The query is assigned the name EG17. The input for this query is

```
EG17
K1 = SCN116/    K2 = NCN=1/    K3 = FG36R/
KEYS = K1 K2 K3 $
END
```

I. QUESTION: What N.N- dialkylphenothiazine carbofamides, with or without additional acyclic ring substituents, are on file?

Structural Representation	Molecular Formula Specifications
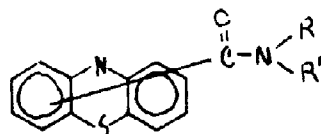
R,R' = any alkyl, same or not
additional substituents on ring allowed

II. ENCODED QUERY:

Query name: *EG17*

Keydefinitions: (Section III.)

KEYS = K1 K2 K3 $

FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

III. KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	KI	SCN IIb	one
Acyclic-Cyclic (A-C)			AN.
Extracyclic (EC)			one
Number of Cyclic Nuclei (NCN)	K2.	NCN-1	
Cyclic Nuclei: non-H Attmts. (DACN)			AN?
Generic Cyclic Nuclei (GCN)	K3	FG 36 R	one
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

148

EXAMPLE 18

Question

Retrieve all metal and metal hydrogen citrates.

Comment

The question involves a family of anions derived from a single organic acid. Note that basic citrates are not sought.

Strategy

The specific functional group key FG94 tags both the carboxylic acid group -COOH and the carboxylate anion -COO⁻ when either of these fragments is attached to an acyclic carbon atom. Thus all metal and metal-hydrogen citrates are assigned this key exactly three times.

In addition to the three carboxylate groups, true responses also contain a secondary hydroxyl (-OH) group, which CIDS tags with the specific functional group key FG82.

$$
\begin{array}{cc}
\underset{\text{---C-O}}{\overset{\displaystyle\overset{O}{\|}}{}} & \underset{\text{O-C---}}{\overset{\displaystyle |}{}} \\[2ex]
\text{FG94} & \text{FG82}
\end{array}
$$

Retrievals are also required to have been assigned the general metal cation key, for which the CIDS code is CN. This key is assigned to each compound which contains either a bare metal cation or the ammonium cation NH_4^+.

Since all of the required compounds are totally acyclic, the query demands responses to have been assigned the key A-C=0, which tags each compound having zero rings. This A-C key is used rather than the key NCN=0 since the latter tags structures and is assigned not only to totally acyclic compounds, but also to compounds that contain both an acyclic structure and a cyclic structure.

Since all responses to the cation key CN must contain either a metal ion or the "metal-like" ammonium ion, it is not necessary to use the general metal key MF M to require a metal in the Hill molecular formula. No molecular formula information is specified for carbon or oxygen either, since the counts of these elements depend on the valence of the metal atom and the charge on the citrate ion. Thus the monosodium salt has the formula $C_6H_7NaO_7$, while the calcium salt structured

149
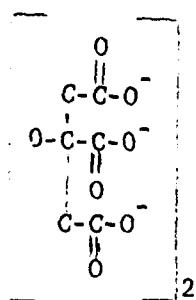
$$
\left[\begin{array}{c}
O \\
\| \\
C-C-O^{-} \\
| \\
O-C-C-O^{-} \\
\| \\
O \\
C-C-O^{-} \\
\| \\
O
\end{array}\right]_{2} \cdot 3Ca^{++}
$$

has the formula $C_{12}H_{10}Ca_3O'_{14}$.

At some stage in processing this query the system must identify those compounds which have been assigned the carboxylate key three times. Since the number of tricarboxylates is expected to be relatively small, it is more efficient to identify these compounds first and then intersect this list with the remaining keys as required in the query. Therefore the carboxylate key is assigned the lowest numbered keydesignation K1, with the remaining keydesignations assigned in expected order of increasing list length as usual.

Although an A/A search for the citrate anion could be conducted to guarantee that all retrievals are true answers, the combination of keys that retrievals must satisfy is considered sufficiently restrictive as to make A/A search unnecessary.

Encoded Query

The query is assigned the name EG18. The input for this query is

```
EG18
K1 = FG94/   K2 = FG82/   K3 = CN/   K4 = A-C=∅/ $
KEYS = 3K1 K2 K3 K4 $
END
```

Query name: EG18

I. QUESTION: *Retrieve all metal and metal-hydrogen citrates.*
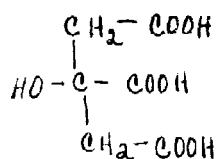
Structural Representation	Molecular Formula Specifications
C H₂ - COOH | HO - C - COOH | CH₂ - COOH *one or more H's of — COOH replaced by metal*	

II. ENCODED QUERY:

Query name: EG18

Keydefinitions: (Section III.)

KEYS = 3K1 K2 K3 K4 $

FORMULA

DEFINE STRUCTURE (Section V.)

STRUCTURE =

END

151

III.

KEYS			
User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF) Acyclic-Cyclic (A-C) Extracyclic (EC)	K1	F-6-94	three AND
Number of Cyclic Nuclei (NCN)	K2	F-6-5-2	one AND
Cyclic Nuclei: non-H Attmts. (DACN)			
Generic Cyclic Nuclei (GCN)	-3	CN	one
Specific Cyclic Nuclei (SCN)			
Specific Functional Group (FG)	K4	A-C=0	one
Nonspec. Diatomics (ND) Nonspec. Monatomics (NM) Hydrocarb. Radicals (HR) Inorganic (IN) Metal Cation (CN) Inorganic Anion (AN) Abnormal Mass (MASS) General Metal (MF M) Nonstructural (DATA) Registry Number (RN)			

IV.

MOLECULAR FORMULA STATEMENT						
Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS

152

EXAMPLE   19

Question

Retrieve all dialkyldimethylammonium halides.

Comment

Besides the structural features illustrated in the question, true responses also have in common a number of molecular formula characteristics which are made use of in this query.

Strategy

All retrievals must contain the following structural features:

(1)   exactly one acylic positively charged nitrogen atom, which is tagged in CIDS by the specific functional group key FG147;

(2)   at least two methyl groups attached to a heteroatom (but possibly more than two, since R and R' can be any alkyl);   the hydrocarbon radical key for methyl attached to a heteroatom is HR1E;

(3)   exactly one halide anion.   CIDS employs the key AN for all inorganic anions (halogen or otherwise).   The key is assigned once to each compound that contains any number or type(s) of inorganic anion(s).

$$\left[\sim N^{+} \right]$$

FG147

C-El~
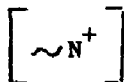
HR1E

The Hill molecular formula of each true answer contains exactly one atom of N and one atom of halogen.   These requirements could be stated by demanding that all retrievals have been assigned the MF N 1 key plus one of the keys MF F 1, MF CL 1, MF BR 1 or MF I 1.   However, all of these keys are expected to have very long keylists.   In view of the selectivity of the structural fragment keys being employed, stipulation of the molecular formula requirements in the formula statement is preferable to using molecular formula keys.

The keydesignations K1, K2 and K3 are assigned to the keys in expected order of increasing list length.

The molecular formula statement proves extremely useful in this query. All retrievals must contain the exact set of elements C, H, N and halogen. By using the general halogen symbol X, the RESTRICTED option can be employed to require that only these elements be present. Nitrogen and halogen are each required to have an exact count of one.

The CONSTRAINTS option is also exercised in this query, since for all true responses the relationship between the carbon count and the hydrogen count is given by $H = 2 * C + 4$.

Encoded Query

The query is assigned the name EG19. The input for this query is

```
EG19
K1 = FG147/
K2 = HR1E/
K3 = AN/
KEYS = K1 2K2 K3 $
FORMULA HILL / RESTRICTED C H N(1) X(1) CONSTRAINTS H = 2 * C + 4/$
END
```

154

Query name: *EG-19*

I. **QUESTION:** *Retrieve all dialkyldimethylammonium halides.*

| Structural Representation | Molecular Formula Specifications |
|---|---|
| $$\left[\begin{array}{c} C \\ | \\ C-N \doteq R' \\ | \\ R \end{array}\right] \cdot X^-$$  $R, R' =$ any alkyl<br>$X = F, Cl, Br, I$ | |
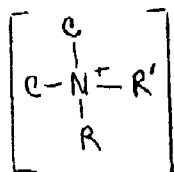
II. **ENCODED QUERY:**

Query name: *EG-19*

Keydefinitions: (Section III.)

KEYS = $K1$  $2K2$  $K3$  $\$$

FORMULA HILL / RESTRICTED C H N(I) X(I) CONSTRAINTS
$$H = 2 * C + 4 \; / \; \$$$

DEFINE STRUCTURE     (Section V.)

STRUCTURE =

END

III. | KEYS

| User's Checklist of Key Types | Key designation | CIDS code | Required assignmt. |
|---|---|---|---|
| Mol. Formula (MF) | K1 | FG147 | One |
| Acyclic-Cyclic (A-C) | | | AN |
| Extracyclic (EC) | | | |
| Number of Cyclic Nuclei (NCN) | K2 | HRIC | OIC |
| Cyclic Nuclei: non-H Attmts. (DACN) | | | AND |
| Generic Cyclic Nuclei (GCN) | K3 | AN | OIC |
| Specific Cyclic Nuclei (SCN) | | | |
| Specific Functional Group (FG) | | | |
| Nonspec. Diatomics (ND) | | | |
| Nonspec. Monatomics (NM) | | | |
| Hydrocarb. Radicals (HR) | | | |
| Inorganic (IN) | | | |
| Metal Cation (CN) | | | |
| Inorganic Anion (AN) | | | |
| Abnormal Mass (MASS) | | | |
| General Metal (MF M) | | | |
| Nonstructural (DATA) | | | |
| Registry Number (RN) | | | |

IV. | MOLECULAR FORMULA STATEMENT

| Formula type | RESTRICTED | Element symbol | Exact count | Lower bound | Upper bound | CONSTRAINTS |
|---|---|---|---|---|---|---|
| HILL | yes | C H N X | 1 1 | | | $H = 2 \times 2 + 4$ |
| | | | | | | |

156

EXAMPLE 20

## Question

What unsubstituted aminocarboxamides are on file that contain from two to eight carbon atoms and any specified nuclide of nitrogen.

## Comment

The user is searching for a specific family of isotopically tagged compounds.

## Strategy

Every response to this query is required to have been assigned the following structural fragment keys:

(a)  the specific functional group key FG34, which tags the unsubstituted carboxamide residue attached to an acyclic carbon atom;

(b)  the specific functional group key FG143, which is assigned to each primary amino group attached to an acyclic carbon atom;

(c)  the abnormal mass key MASS which is assigned to every compound that contains one or more "abnormal" isotopes of any element(s);

(d)  the key A-C=O which is assigned to every compound that has a total of zero rings.  This A-C key is used rather than the zero cyclic nuclei key (NCN=0) since the NCN=0 key tags structures, and its keylist therefore references not only totally acyclic compounds but also compounds that contain both an acyclic and a cyclic structure.

$$\overset{\overset{\displaystyle O}{\overset{\|}{}}}{- - - C - N} \qquad\qquad \qquad — N$$

FG34                                     FG143

Since the Hill molecular formula of every true answer contains one O and two N, the quantitative molecular formula keys MF O 1 and MF N 2 could be demanded.  However, since both of these MF keys have long keylists and since the structural fragment keys used in the query are reasonably restrictive, it is more efficient to specify the molecular formula requirements in the formula statement.

By specifying the exact counts of N and O in the Hill formula and limiting the element types to C, H, N and O with the RESTRICTED option, the formula

157

statement insures that the only heteroatom-containing functional groups in retrievals are a single amino group and a single carboxamide group. The range of two to eight for the carbon count is specified; since unsaturation is permitted, no limitations on the hydrogen count are included.

While the MASS key insures that responses contain a specified nuclide of some element(s), an A/A search is required to retrieve only those which contain a nuclide of nitrogen. Conducting an A/A search for the fragment

$$^{n}N\sim$$
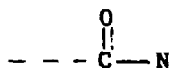
in which the nitrogen atom has the indefinite mass number $\underline{n}$ retrieves the required N-labeled compounds.

## Encoded Query

The query is assigned the name EG20. The structurespecification of the fragment to be A/A searched contains an abnormality string that requires the nitrogen to have any "abnormal" mass. The input for this query is

```
EG20
K1 = MASS/  K2 = FG34/  K3 = FG143/  K4 = A-C=0/
KEYS = K1  K2  K3  K4 $
FORMULA HILL / RESTRICTED C(2,8) H N(2) O(1) / $
DEFINE STRUCTURE S1 = /1N80. (M1=N.)/ $
STRUCTURE = S1 $
END
```

Query name: *EG20*

I. QUESTION: *What unsubstituted aminocarboxamides are on file that contain from 2 to 8 C atoms and any specified nuclide of nitrogen?*
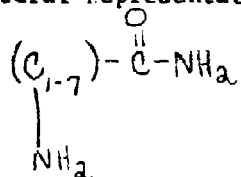
Structural Representation	Molecular Formula Specifications
$(C_{1-7})-\overset{\overset{O}{\|\|}}{C}-NH_2$ *either or both N's abnormal mass* NH_2	

II. ENCODED QUERY:

Query name: *EG20*

Keydefinitions: (Section III.)

KEYS = *K1 K2 K3 K4 $*

FORMULA *HILL / RESTRICTED C(2,8) H N(2) O(1) / $*

DEFINE STRUCTURE (Section V.)

STRUCTURE = *S1 $*

END

159

II. | KEYS

User's Checklist of Key Types	Key designation	CIDS code	Required assignmt.
Mol. Formula (MF)	$K1$	$MASS$	one
Acyclic-Cyclic (A-C)			AND
Extracyclic (EC)			
Number of Cyclic Nuclei (NCN)	$K2$	$FG34$	one
Cyclic Nuclei: non-H Attmts. (DACN)	$K3$	$FG143$	AND
			one
Generic Cyclic Nuclei (GCN)			
Specific Cyclic Nuclei (SCN)	$K4$	$A-C = \emptyset$	AND
			one
Specific Functional Group (FG)			
Nonspec. Diatomics (ND)			
Nonspec. Monatomics (NM)			
Hydrocarb. Radicals (HR)			
Inorganic (IN)			
Metal Cation (CN)			
Inorganic Anion (AN)			
Abnormal Mass (MASS)			
General Metal (MF M)			
Nonstructural (DATA)			
Registry Number (RN)			

IV. | MOLECULAR FORMULA STATEMENT

Formula type	RESTRICTED	Element symbol	Exact count	Lower bound	Upper bound	CONSTRAINTS
HILL	yes	C		2	8	
		H				
		N	2			
		O	1			

160

V. ATOM-BY-ATOM SEARCH

Structure(s):

$^{m}N \sim$

Structure name	Structurespecification	No. of Occurrences
S1	1N8Ø. (M1 = m.)	one

LITERATURE CITED

1. B. Suerr, The CIDS Multi-Terminal Command Language for Teletypes University of Pennsylvania, Philadelphia, Pa., October 1968

2. P. R. Weinberg, A Guide to the CIDS Retrieval Language, University of Pennsylvania, Philadelphia, Pa., November 1967

3. C. T. Van Meter, E. N. Goldschmidt, M. Milne, Handbook of CIDS Chemical Search Components, CIDS No. 6 Status Report, University of Pennsylvania, Philadelphia, Pa., December 1968

4. R. V. Powers, Querying a Real Time Chemical Information Retrieval System. University of Pennsylvania, Philadelphia, Pa., May 1969

5. M. Milne, Chemical Editing Conventions I, University of Pennsylvania, Philadelphia, Pa., August 1968

APPENDIX A

THE CIDS MULTITERMINAL COMMAND LANGUAGE FOR TELETYPES

This appendix is intended to serve as a user's guide for teletypes (TTY's) in the CIDS multiterminal retrieval system. The teletypes communicate directly with a PDP-8 computer, the primary purpose of which is the accumulation of input and the distribution of output for each of the various terminals. The PDP-8 in turn communicates with an IBM 7040 system that actually conducts the search. This appendix describes the language for inputting and editing queries via the PDP-8 and for carrying through with search and retrieval in the 7040.

Introduction

Each line that is typed at a system TTY is interpreted either as a command or as text depending on the first character in the line. If this character is '@', (the <u>ALT MODE</u> or <u>ESC</u> key, whichever is present on the TTY being used), the line is interpreted as a command and is executed. If the first character is not '@', the line is interpreted as text, and is either transmitted to the 7040 for immediate processing (by terminating the line with an exclamation point (!)) or else stored by the PDP-8 for future processing (by terminating the line with a 'carriage return').

The commands that may be issued fall into two categories: those used in editing lines of stored text either to modify query demands or to correct errors; and those used to carry out processing of the query by the 7040.

Lines of text that are stored are assigned line numbers which may be used in subsequent commands to identify the line(s) to be acted on. The number to be assigned to the next line of text that is stored (called the <u>current line number</u>) is automatically printed out at the beginning of each line. Each time a line of text is stored, the current line number is incremented. The line number is not incremented following command lines or text lines that are not stored, and the numbers that precede such lines on the TTY printout have no significance.

A few of the characters on the TTY keyboard, such as the '@', '! ' and carriage return mentioned above, have special meaning in the TTY language. An outline of these special characters and their meanings appears below. Following the outline, the various teletype commands are described, and a number of examples of their use are provided.

165

Special Characters

The special characters that have particular meanings in the teletype language are:

Printed Character	Name	Meaning
	carriage return	In a command line: End of command. In a text line: Store this line and start a new one.
↑	up-arrow	Ignore this line.
←	back-arrow	Ignore the last character typed. The character that is ignored is printed out.
@	'ALT MODE' or 'ESC'	At the beginning of a line: Interpret this line as a command. Elsewhere in the line, no special meaning.
"	ditto	Copy the character in the corresponding position in the last line typed in at this console.
	Control C (CTRL and C are struck simultaneously)	Copy the character in the corresponding position in the last line printed out at this console.
!	exclamation point	Terminating a text line: Send this line of text to the 7040 for immediate processing; do not store it in the PDP-8, or assign it a line number.

Commands

The commands employed in the teletype language can be subdivided into two categories: commands used in editing stored text, and commands that effect interaction between the PDP-8 and the 7040.

I. Editing Commands

    a) @LINE prints the current line number at the beginning of the next line on the teletyped page.

    b) @PRINT prints specified successive lines of text from the user's file (i.e., successive lines of stored text) onto the teletype. This command is structured as follows:

        @PRINT n   causes the single line n to be printed.

        @PRINT n,m   causes lines n through m to be printed out.

        @PRINT n,   causes line n and all succeeding lines to be printed out.

    c) @DELETE cause a line or a group of lines of stored text to be deleted, and is structured similarly to @PRINT.

    d) @INSERT n causes the value of the current line number to become n. Lines that are typed following the INSERT command are assigned successive line numbers beginning with number n. Insertion continues until the APPEND command is typed, which causes the remaining originally stored lines to be renumbered to follow the inserted lines, and causes the current line number to return to its usual value (i.e., one greater than the total number of lines currently stored).

    e) @ALTER combines the effect of the DELETE and INSERT commands. As with insertion, alteration is terminated by the APPEND command. The arguments given are the same as those for the DELETE command as follows:

        @ALTER m,n   causes lines m through n to be deleted and a line (or lines) to be inserted starting with line number m.

        @ALTER m   deletes line m and inserts a line (or lines) starting with line number m.

167

@ALTER m, deletes all lines starting with line m and
inserts beginning with line m.

f) **@APPEND** terminates insertion resulting from a previously issued
INSERT or ALTER command.

g) **@MOVE m,n,p** causes lines m through n inclusive to be moved
to follow line p. The lines that originally followed immediately
after line p are renumbered accordingly.

II. **Commands Resulting in Interaction with 7040**

The following commands result in some interaction between the PDP-8
and the 7040.

a) **@INIT** initializes the flags, buffers and input file for the
console at which it is typed. Any previously saved text in
the PDP-8 is destroyed. If the 7040 is connected the CLEAR
command is also executed as part of the INIT command, terminat-
ing any search that may be going on for this console and des-
troying all information related to a terminated search.

b) **@PROCESS n,m** causes lines n through m to be sent to the 7040.
If the last line sent is an "END" statement, the 7040 will
consider the lines it receives as a complete query and pre-
process it. As a result, the PDP-8 will get accession list
information and/or error messages as output.

c) **@START** signals the 7040 to begin applying any formula statement
and A/A search that are present to the accession list entries.
As answers are found they are sent to the PDP-8, which routes
them back to be output by the appropriate device. Unless com-
mands require otherwise, the complete record is punched by the
TTY on paper tape to be printed out on a Dura Mach chemical
typewriter.

d) **@STOP** results in a temporary halting of the actual search pro-
cess. The accumulated output continues to be sent to the console.
Search can be continued by means of the START command.

168

e) @CLEAR stops the search in the 7040 and destroys all information in the 7040 related to that particular query. The search can be restarted after reissuing the PROCESS command. The system responds with the message (SEARCH RECORDS CLEARED). The CLEAR command differs from the INIT command in that it has no effect on the user's input file in the PDP-8.

f) Three output option commands may be typed before @START. If none of these appear, the output for that query will include the entire chemical record and will be sent to the remote console in Dura paper tape code. The following commands may be typed in any order but must be typed before @START.

1) @OUTPUT REGISTRY requests only registry numbers for answers to a given query. These will be typed on the TTY unless @OUTPUT PRINTER is also specified, in which case the answers will be sent only to the line printer.

2) @OUTPUT STAT requests only statistics, i.e., the search time and the total number of actual answers. These will appear on the TTY. If @OUTPUT PRINTER is also specified, the statistics will also appear on the line printer.

3) @OUTPUT PRINTER specifies that all query answers for that console go to the chemical line printer.

These options must be repeated for each query where they are desired and must be typed before @START.

III. The ECHO Command

The ECHO command enables the PDP-8 operator to monitor the activity at another remote TTY and thus to receive messages from that TTY. This command may be issued from the PDP-8 console TTY only, and is structured as follows:

@ECHO n   causes everything on TTY n, both input and output, to be printed out on the PDP-8 TTY as well. n is the number (called the console number) being used to represent a particular terminal during the current querying session. This number is automatically assigned by the system as each terminal dials in, and is printed out in the response to a PROCESS command immediately before the query name.

169

While in ECHO mode, the PDP-8 TTY can also send messages to the TTY being echoed, as follows:

If the CTRL and BELL keys are struck simultaneously on the PDP-8 TTY at the beginning of a new line, a bell is sounded at both teletypes (to signal the operator of the echoed TTY), and the line that is then typed at the PDP-8 TTY is also printed out on the echoed TTY.

Simultaneously depressing CTRL and BELL at the echoed TTY also sounds a bell at both teletypes, and can be used to alert the PDP-8 operator that a message is about to be echoed. Such message lines should be terminated with up-arrow ↑ to cause these lines to be ignored and not saved as input text for that console.

IV. Expansion of Logical Expressions: Use of OUTPUT LOGIC

To insure that parentheses in KEYS and STRUCTURE logical statements are properly located, a user can require the search system to expand these statements by typing the words OUTPUT LOGIC immediately after the query name. This causes the expanded version of every logical statement in the query to be included in the search system's response to the PROCESS command (see Example 13 following).

Example 1: INIT, PRINT Commands

```
        @INIT
PROJECT CIDS
(SEARCH RECORDS CLEARED)                  ⎫
0001 CIDS                                 ⎪
0002 K1 = FG143R/                         ⎪
0003 K2 = FG35R/                          ⎪
0004 K3 = SCN48/                          ⎬  2
0005 KEYS = K1 K2 K3 $                     ⎪
0006 FORMULA HILL / C(12, ) / $           ⎪
0007 END                                  ⎭
0008 @PRINT 4                     ___  3
0004 K3 = SCN48/
0008 @PRINT 4,                    ___  4
0004 K3 = SCN48/
0005 KEYS = K1 K2 K3 $
0006 FORMULA HILL / C(12, ) / $
0007 END
0008 @PRINT 4, 6                  ___  5
0004 K3 = SCN48/
0005 KEYS = K1 K2 K3 $
0006 FORMULA HILL / C(12, ) / $
0008
```

(1)   The command @INIT is issued. All previously stored text for this console
       is destroyed; PROJECT CIDS, (SEARCH RECORDS CLEARED), and the current
       line number 0001 (with leading zeros to four places) are printed out.

(2)   Seven lines of query text are input and stored. The current line number
       is incremented after each line stored.

(3)   A command to print the fourth line of stored text is typed, followed by
       a carriage return. The line number, 0004 and the text of that line are
       printed out, followed by the current line number. Note that the line
       number is _not_ incremented following the @PRINT command.

(4)   A command to print line 4 and all subsequent lines is issued, followed
       by a carriage return. The appropriate lines are printed out, followed
       by the current line number (still 0008).

(5)   A command to print lines 4 through 6 inclusive is issued, followed by
       a carriage return. The indicated lines are output, followed by the
       current line number (still 0008).

171

Example 2: PROCESS, START Commands

```
0001 CIDS
0002 K1 = FG35R/    K2 = SCN46/    K3 = FG143R/
0003 KEYS = K1 K2 K3 S
0004 FORMULA HILL / (17, )/S
0005 END
0006 EA20
0007 K1 = FG112/
0008 K2 = FG113/
0009 K3 = FG114
0010 KEYS = K1 OR K2 OR K3 S
0011 END
0012 @PROCESS 1,S
  1013 HRS. 700416
        COMMENTS FOR QUERY        2CIDS
        000036 ITEMS IN ACC. LIST FOR THIS QUERY
@START
0012 DNEJAFN+@EJLTMQJF\QQUVGFGQUVLPAVQFYU FF
ACTIVE SEARCH TIME
           000 MINUTES
            08 SECONDS
            05 60THS
  0008 ANSWERS FOR THIS QUERY

************ READY FOR A NEW QUERY ************
```

1 (brackets lines 0001-0011)
2 (line 0012 @PROCESS 1,S)
3 (brackets 1013 HRS through ITEMS)
4 (@START)
5 (line 0012 DNEJ...)
6 (brackets ACTIVE SEARCH TIME through READY FOR A NEW QUERY)

(1) Two queries containing a total of eleven lines of text are input and stored.

(2) A command to process the first query (lines 1-5) is issued.

(3) The system identifies the date and time, the query name, and the number of compounds that satisfy the keys. The number that precedes the query name 'CIDS' is the console number, by which this TTY is being represented during the current querying session. (This number is assigned by the system as each terminal dials in.)

(4) The START command causes the formula statement (and the A/A search when one is present) to be applied to the responses to the keys; compounds that successfully pass these tests are output.

(5) Answers are being punched on paper tape to be printed out on a DURA MACH chemical typewriter. (Transmission of Dura code to the TTY punch causes spurious characters to be printed out on the teletype.)

(6) Search statistics — search time and total number of answers — are output. The system signals that a new query may now be processed.

172

Example 3: INSERT, APPEND Commands

```
0001 CIDS                                        ⎫
0002 K1 = FG143R/                                 ⎪
0003 KEYS = K1 K2 K3 $                            ⎬  1
0004 FORMULA HILL / C(17, ) / $                   ⎪
0005 END                                         ⎭
0006 @INSERT 3                             —— 2
0003 K2 = FG35R/                                  ⎫  3
0004 K3 = SCN48/                                  ⎭
0005 @APPEND                                —— 4
0008 @PRINT 1,                                    ⎫
0001 CIDS                                         ⎪
0002 K1 = FG143R/                                 ⎪
0003 K2 = FG35R/                                  ⎪
0004 K3 = SCN48/                                  ⎬  5
0005 KEYS = K1 K2 K3 $                            ⎪
0006 FORMULA HILL /C(17, ) / $                    ⎪
0007 END                                         ⎭
0008
```

(1) In the five line query that has been stored, keydefinitions for keys K2 and K3 (referenced in line 3) have been omitted, and are to be inserted.

(2) A command is issued to insert, immediately preceding line 3, all lines between the INSERT command and a subsequent APPEND command. The INSERT command causes the value of the current line number to become 0003.

(3) Two lines of text are input for insertion preceding the original line 3.

(4) The APPEND command is issued to terminate insertion.

(5) The entire block of stored text is printed out, illustrating the results of the insertion. Note that the lines originally numbered 3 through 5 are renumbered to follow the inserted lines. The current line number is returned to its usual value, i.e., one greater than the number of lines currently stored.

173

```
0001 CIDS                              ⎫
0002 K1 = FG143R/                      ⎪
0003 K2 = FGE5R/                       ⎪
0004 K2 = FG35R/                       ⎪
0005 K4 =                              ⎪
0006 K3 = SCN 38/                      ⎬   1
0007 K3 = SCN 48/                      ⎪
0008 KEYS = K1 K2 K3 $                 ⎪
0009 FORMULA HILL / C(17, ) / $        ⎪
0010 END                              ⎭
0011 @DELETE 3                         ⎫
0010 @PRINT 1,                         ⎪
0001 CIDS                              ⎪
0002 K1 = FG143R/                      ⎪
0003 K2 = FG35R/                       ⎪
0004 K4 =                              ⎬   2
0005 K3 = SCN 38/                      ⎪
0006 K3 = SCN 48/                      ⎪
0007 KEYS = K1 K2 K3 $                 ⎪
0008 FORMULA HILL / C(17, ) / $        ⎪
0009 END                              ⎭
0010 @DELETE 4,5                       ⎫
0008 @PRINT 1,                         ⎪
0001 CIDS                              ⎪
0002 K1 = FG143R/                      ⎪
0003 K2 = FG35R/                       ⎬   3
0004 K3 = SCN 48/                      ⎪
0005 KEYS = K1 K2 K3 $                 ⎪
0006 FORMULA HILL / C(17, ) / $        ⎪
0007 END                              ⎭
0008
```

(1) While inputting this query, typing errors (bracketed) in line 3 and in
     lines 5 and 6 are corrected in line 4 and line 7 respectively. The
     three incorrect lines 3, 5 and 6 are to be deleted.

(2) Commands are issued to delete line 3 and print the remaining text. Note
     that the lines originally numbered 4-10 are renumbered 3-9 to fill in
     for the deleted line.

(3) A command is issued to delete lines 4 through 5 inclusive (the two remain-
     ing incorrect lines, renumbered following deletion of line 3). The
     resulting text is printed out.

Example 5: ALTER, APPEND Commands

```
0001 CIDS
0002 K1 = FH143R/                    ⌐
0003 K2 = FG35                       │
0004 K2 = HG35R/                     ⌐  1
0005 K3 = SW48/                      │
0006 KEYS = K1 K2K3 $                ┘
0007 FORMULA HILL / C(17, )/$
0008 END
0009 @ALTER 2,5                      ⌐  2
0002 K1 = FG143R/                    ⌐
0003 K2 = FG35R/                     │  3
0004 K3 = SCN48/                     ┘
0005 @APPEND
0008 @PRINT 1,                       ⌐
0001 CIDS                            │
0002 K1 = FG143R/                    │
0003 K2 = FG35R/                     │
0004 K3 = SCN48/                     ┘  4
0005 KEYS = K1 K2 K3 $               │
0006 FORMULA HILL / C(17, )/$        │
0007 END                            ┘
```

(1) The user requires this query to include the three keydefinitions
K1 = FG143R/, K2 = FG35R/, and K3 = SCN48/. Therefore lines 2 through
5 are all incorrect and must be replaced.

(2) The command @ALTER 2,5 deletes lines 2 through 5 inclusive and causes
all lines that are input between the ALTER command and a subsequent
APPEND command to be inserted beginning at line 2. The number of lines
inserted may be greater than, less than or equal to the number of lines
deleted; the remaining lines of stored text will be renumbered as required.

(3) The correct lines are typed in, followed by the APPEND command.

(4) The complete stored text is printed out, illustrating required renumber-
ing of the lines of text following the altered lines.

Example 6:   STOP. CLEAR Commands

```
0001 CIDS                                                        ⌉
0002 K1 = FG143R/                                                │
0003 K2 = FG35R/                                                 │
0004 K3 = SCN48/                                                 │
0005 KEYS = K1 K2 K3 $                                           │
0006 FORMULA HILL / C(12,) /$                                    ⎬ 1
0007 END                                                         │
0008 @PROCESS 1,7                                                │
  1344 HRS. 700323                                               │
      COMMENTS FOR QUERY        1CIDS                            │
      000027 ITEMS IN ACC. LIST FOR THIS QUERY                  ⌋
@START                                            ─ 2
0008 DNEJ AFN←@EJLTMQJF\ Q0GFGQPTL VAYFVYU YLY          UU T          ⎫
0008 @STOP                                                           ⎬ 3
0008  FLLLLY                        PYY                            ⎭
0008 @CLEAR                                        ─ 4
0008 (SEARCH RECORDS CLEARED)
```

(1) A seven line query is input and processed.

(2) The START command causes the formula statement and A/A search (when present) to be applied to the responses to the keys. Compounds that pass are punched in Dura code.

(3) @STOP causes the search to be halted, but output of compounds that have already passed continues. Search can be resumed by reissuing @START.

(4) @CLEAR causes both search and output for this query to cease, and all record of which compounds are waiting to be tested or output to be destroyed. The system responds with (SEARCH RECORDS CLEARED). The complete set of responses can be obtained only by rerunning the entire query.

176

Example 7: Exclamation Point (!), LINE Command.

```
0001 CIDS!                                                          ⌉
0001 K1 = FG143R/!                                                  |
0001 K2 = FG35R/!                                                   |
0001 K3 = SCN48/!                                                   |
0001 KEYS = K1 K2 K3 $!                                             ⎬  1
0001 FORMULA HILL / C(22, ) /$!                                     |
0001 END!                                                           |
0001  1117 HRS. 700416                                              |
      COMMENTS FOR QUERY        ICIDS                               |
      000036 ITEMS IN ACC. LIST FOR THIS QUERY ⌋
@LINE                                                               ⌉
0001                                                                ⎬  2
```

(1)   Seven lines of query text are typed. Termination of each line with an
      exclamation point (rather than a carriage return) causes each line to
      be sent to the 7040 for processing. Since these lines are not stored,
      no line numbers are assigned. As soon as the END statement is received,
      processing of the keys is carried out automatically. (Therefore, no
      PROCESS command is required). The START command can now be issued to
      complete the processing of this query and output responses.

(2)   The LINE command causes the current line number to be printed out.
      Since the preceding text has not been stored, the current line number
      is still 1.

      Queries that are sent to be processed immediately as above cannot be
printed out, altered or restarted except by retyping the entire query.

177

Example 8: Ditto ("), CTRL C, Up-Arrow (↑) and Back-Arrow (←)

```
0001 CIDS                                              ⎫
0002 K1 = FG143R/                                      ⎬ 1
0003 K2 = FT35R/↑                                      ⎭
0003 K2 = FG35R/                                       --- 2
0004 K3 = SCN48/
0005 KEYS = K1 K2 K4 $
0006 FOXYYXRMULA HILL / C(22, )/$                      -- 3
0007 END
0008 @PRINT 1,
0001 CIDS                                              ⎫
0002 K1 = FG143R/                                      |
0003 K2 = FG35R/                                       | 4
0004 K3 = SCN48/                                       |
0005 KEYS = K1 K2 K4 $                                 |
0006 FORMULA HILL / C(22, )/$                          |
0007 END                                               ⎭
0008 @PRINT 5                                          ⎫
0005 KEYS = K1 K2 K4 $                                 |
0008 @ALTER 5                                          ⎬ 5
0005 KEYS = K1 K2 K3, $                                |
0006 @APPEND                                           ⎭
0008 @PRINT 1,                                         ⎫
0001 CIDS                                              |
0002 K1 = FG143R/                                      |
0003 K2 = FG35R/                                       | 6
0004 K3 = SCN48/                                       |
0005 KEYS = K1 K2 K3 $                                 |
0006 FORMULA HILL / C(22, )/$                          |
0007 END                                               ⎭
0008 @START↑                                           ⎫
0008 @PROSSCDSSSSDESS 1,                               |
 1125 HRS. 700416                                      ⎬ 7
     COMMENTS FOR QUERY       1CIDS                    |
     000036 ITEMS IN ACC. LIST FOR THIS QUERY          ⎭
```

(1) A typing error, a "T" instead of a "G" in the CIDS code, is detected in line 3 after the slash has been typed. The line is terminated with up-arrow rather than carriage return, causing the entire line to be ignored and current line number to remain unchanged.

(2) The corrected version of line 3 is input; each of the bracketed characters was copied from the previously typed line (i.e., the deleted version of line 3) by striking ditto (").

178

(3) The characters 'XY' incorectly typed in the formula statement (line 6) are deleted using back-arrow (←) as follows:

Typing error made

    Back-arrow struck to delete last character. The character
       being deleted (Y) prints out.

        Back-arrow again used to delete last character (now 'X'
             since the 'Y' is gone). The deleted 'X' prints out.

006 FOXYYXRMULA HILL / C(22,)/$

(4) The entire query is printed out to illustrate the corrections.

(5) An error in line 5 is detected: 'K4' should read 'K3'. The incorrect line 5 is printed out and a command is given to alter 5. The correct version of line 5 is input by copying the correct parts (bracketed above) of the last line printed out, (i.e., the incorrect line 5), depressing the CTRL key and striking the character C for each character copied.

(6) The entire query is printed out for examination.

(7) The use of up-arrow and back-arrow on commands (rather than on text) is permitted.

Example 9:   OUTPUT REGISTRY Command

```
0001 CIDS                                              ⎤
0002 K1 = FG143R/                                      |
0003 K2 = FG35R/                                       |
0004 K3 = SCN 48/                                      |
0005 KEYS = K1 K2 K3 $                                 |
0006 FORMULA HILL / C(22, )/$                          ⎬ 1
0007 END                                               |
0008 @PROCESS 1,                                       |
 1127 HRS. 700416                                      |
     COMMENTS FOR QUERY       1CIDS                    |
     000036 ITEMS IN ACC. LIST FOR THIS QUERY          ⎦
@OUTPUT REGISTRY                                       ⎤
0008 @START                                            |
0008       QUERY NUMBER CIDS                           |
 RN   A0106502                                         ⎬ 2
        TN   T706151                                   |
 RN   A0110457                                         |
        TN   T711452                                   ⎦

ACTIVE SEARCH TIME                                     ⎤
        000 MINUTES                                    |
        05 SECONDS                                     |
        20 60THS                                       ⎬ 3
 0002 ANSWERS FOR THIS QUERY                           |
                                                       |
************ READY FOR A NEW QUERY ************        ⎦
```

(1) A seven line query is stored and processed.

(2) The OUTPUT REGISTRY command preceding the START command causes the master
     registry number and all local identification numbers of responses to be
     output by the teletype.   No DURA paper tape is punched.

(3) Search statistics are automatically output; the system signals that it
     is ready to process another query.

180

Example 10:   OUTPUT STAT Command

```
0001 CIDS                                                  ⌐
0002 K1 = FG143R/                                          │
0003 K2 = FG35R/                                           │
0004 K3 = SCN4R/                                           │
0005 KEYS = K1 K2 K3 $                                     │
0006 FORMULA HILL / C(22, )/ $                            ⟩ 1
0007 END                                                   │
0008 @PROCESS 1,                                           │
  1130 HRS. 700416                                         │
     COMMENTS FOR QUERY        1CIDS                       │
     000036 ITEMS IN ACC. LIST FOR THIS QUERY             ⌐
@OUTPUT STAT                                               │
0008 @START                                                │
0008      QUERY NUMBER CIDS                                │
                                                           │
ACTIVE SEARCH TIME                                         │
        000 MINUTES                                       ⟩ 2
         05 SECONDS                                        │
         19 60THS                                          │
 0002 ANSWERS FOR THIS QUERY                               │
                                                           │
************ READY FOR A NEW QUERY ************            ⌐
```

(1) A seven line query is input and processed.

(2) The OUTPUT STAT command preceding the START command causes the search
     statistics printed out by the teletype to be the only output. To obtain
     the registry numbers or the structures, it is necessary to reissue the
     PROCESS and START commands.

Example 11:   OUTPUT PRINTER Command

```
0001 CIDS
0002 K1 = FG143R/
0003 K2 = FG35R/
0004 K3 = SCN48/
0005 KEYS = K1 K2 K3 $
0006 FORMULA HILL / C(22, )/ $
0007 END
0008 @PROCESS 1,
  1132 HRS. 700416
      COMMENTS FOR QUERY        1CIDS
      000036 ITEMS IN ACC. LIST FOR THIS QUERY
@OUTPUT PRINTER
0008 @START
0008
ACTIVE SEARCH TIME
        000 MINUTES
        07 SECONDS
        59 60THS
 0002 ANSWERS FOR THIS QUERY

*********** READY FOR A NEW QUERY ***********
```

(1) A seven line query is input and processed.

(2) The OUTPUT PRINTER command preceding the START command causes the entire
     record (including nomenclature) for each response to be sent to the
     chemical line printer located at Edgewood Arsenal. No DURA tape is
     punched.

         The OUTPUT PRINTER command can be combined with the OUTPUT REGISTRY
     command to send only the master registry numbers and local identification
     numbers to the printer. Similarly, @OUTPUT PRINTER can be combined with
     @OUTPUT STAT to send only the search statistics to the printer.

Example 12:   MOVE Command

```
0001 CIDS                                                                    ⎫
0002 K1 = FGI43R/                                                            ⎪
0003 K2 = FG35R/                                                             ⎪
0004 K3 = SCN48/                                                             ⎪
0005 KEYS = K1 K2 K3 $                                                       ⎪
0006 DEFINE STRUCTURE                                                        ⎬ 1
0007     S = / 1*C4*2-80-14.    2*C1 3-80.    3N.    4C25-16.                ⎪
0008           50.    6N80./ $                                               ⎪
0009 STRUCTURE = S $                                                         ⎪
0010 FORMULA HILL / C(22, )/$                                                ⎪
0011 END                                                                     ⎭
0012 @MOVE 10,11,5                                                           ⎫
0012 @PRINT 1,                                                               ⎪
0001 CIDS                                                                    ⎪
0002 K1 = FGI43R/                                                            ⎪
0003 K2 = FG35R/                                                             ⎪
0004 K3 = SCN48/                                                             ⎪
0005 KEYS = K1 K2 K3 $                                                       ⎬ 2
0006 FORMULA HILL / C(22, )/$                                                ⎪
0007 END                                                                     ⎪
0008 DEFINE STRUCTURE                                                        ⎪
0009     S = / 1*C4*2-80-14.    2*C1 3-80.    3N.    4C25-16.                ⎪
0010           50.    6N80./ $                                               ⎪
0011 STRUCTURE = S $                                                         ⎭
```

(1)   Eleven lines of query text are input and stored.   The user wishes to
eliminate the DEFINE STRUCTURE statement and the STRUCTURE logical
statement from the query without destroying them.

(2)   The MOVE command is issued to cause lines 10-11 inclusive to be moved
to directly follow line 5.   The resulting stored text is commanded to
be printed out.

183

Example 13: Use of OUTPUT LOGIC


```
    @PRINT 1,5                                              —  1
0001 CIDS        OUTPUT LOGIC
0002 K1 = SCN124/  K2 = SCN125/  K3 = FG96R/
0003 K4 = CN/
0004 KEYS = (K1 OR K2) K3 NOT 2K3 K4 $
0005 END
0013 @PROCESS 1,5
    1609 HRS.  700424
        COMMENTS FOR QUERY          2CIDS
            EXPANDED LOGICAL EXPRESSION      KEYS        2
            K1 K3 NOT 2K3 K4       .
                    OR
            K2 K3 NOT 2K3 K4
        000008 ITEMS IN ACC. LIST FOR THIS QUERY
```


(1) To require the system to output the expanded version of the KEYS logical
     statement in line 4 (as well as any other logical statements present),
     the phrase OUTPUT LOGIC is typed immediately after the query name CIDS.

(2) The PROCESS command is issued; the system's response includes the
     expanded KEYS logical expression.

184

APPENDIX B

CONVENTIONS FOLLOWED IN PROTOTYPES

Listed below are the conventions followed in interpreting the prototypes for the query statements contained in Section 3.5.

(a) Strings of small letters (e.g., keydesignation, structurename, count) are variables for which strings of upper case characters are substituted.

(b) Strings of upper case letters and numbers stand for themselves, e.g., FORMULA, DEFINE STRUCTURE, ADDEND.

(c) Where a choice is involved, braces { } are used to list the alternatives. Default options are indicated by underlining.

(d) Brackets [ ] are used to indicate that their contents are optional.

(e) Ellipses (i.e., ...) are used in the standard manner to mean "continue in the same way."

(f) Subscripting is used in the standard manner to indicate a sequence.

(g) Unless otherwise stated in the description of a particular statement, the following rules govern the use of blanks (space characters) in query statements.

    i   Blanks must not appear within a complete word. (Example: DE FINE STRUCTURE is invalid and must be written DEFINE STRUCTURE).

    ii Blanks are permitted between language elements, e.g., between a complete word and a punctuation mark.

    iii Blanks must be used to separate strings consisting entirely of letters and digits. For example, FORMULAHILL / S / $ is unacceptable because the words FORMULA and HILL run together. However, both FORMULA HILL / S/$ and FORMULA HILL/S/$ are acceptable since / and $ are special characters which can be distinguished from the word HILL and the letter S by the scanning program.

185

Example:

The prototype for the molecular formula statement (Sec. 3.5.2.3) is

$$\left[\text{FORMULA} \left\{ \begin{array}{l} \text{HILL/} \\ \text{ADDEND/} \end{array} \right\} \text{formula/} \quad \text{formula/} \quad ... \quad \text{formula/} \ \$ \right]$$

which is interpreted as follows:

1. The outer brackets indicate that the use of the formula statement in a query is optional.

2. When included, the formula statement consists of

   a. the word 'FORMULA' followed by

   b. either ' HILL / ' or ' ADDEND / ' followed by

   c. one or more terms of the form 'formula /' where formula is as described below, followed by

   d. a dollar sign ($).

Each formula follows the prototype

$$\left[\text{RESTRICTED}\right] \left\{ \begin{array}{l} \text{el} \\ \text{el (n)} \\ \text{el (lb,)} \\ \text{el (,ub)} \\ \text{el (lb,ub)} \end{array} \right\} ... \left\{ \begin{array}{l} \text{el} \\ \text{el (n)} \\ \text{el (lb,)} \\ \text{el (,ub)} \\ \text{el (lb,ub)} \end{array} \right\} \left[\begin{array}{l} \text{CONSTRAINTS} \quad \text{el} = a*\text{el} \pm b ... \\ \qquad\qquad\qquad \text{el} = a*\text{el} \pm b \end{array} \right]$$

which is interpreted as follows:

1. The word 'RESTRICTED' is enclosed in brackets and is therefore optional in each formula.

2. Each formula must contain one or more terms each of which consists of either an element symbol (el) alone, or an element symbol followed by the exact count (n) or the lower bound (lb) and/or the upper bound (ub) for that element in parentheses.

3. The CONSTRAINTS feature is enclosed in brackets and is therefore optional. If included, it consists of the word 'CONSTRAINTS' followed by one or more equations. Each of these equations has the general form $\text{el} = a*\text{el} \pm b$ where the el's are element symbols and a and b are integers. (The limits for a and b are given in Sec. 3.5.2.3.)

186

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
UNIVERSITY OF PENNSYLVANIA Philadelphia, Pennsylvania 19104	UNCLASSIFIED
	2b. GROUP NA

3. REPORT TITLE

CIDS No. 7 QUERY FORMULATION AND ENCODING

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Status Report, December 1967 - April 1970

5. AUTHOR(S) (First name, middle initial, last name)

Milne, Margaret, Weinberg, Paul R.

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
April 1970	192	5

8a. CONTRACT OR GRANT NO. DAAA15-69-C-0140	9a. ORIGINATOR'S REPORT NUMBER(S)
b. PROJECT NO.	STARE / CIDS No. 7
c. Task: 2P062101A72702	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT This document is subject to special export controls and each transmittal to a foreign national or a foreign government may be made only with prior approval of the Commanding Officer, Edgewood Arsenal, ATTN: SMUEA-TSTI-T, Edgewood Arsenal, Maryland 21010

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
Army Chemical Data Systems	Edgewood Arsenal Technical Support Directorate, Edgewood Arsenal, Maryland 21010 (Stanley Goldberg, Proj. O., 671-2807)

13. ABSTRACT >This document is intended to serve as a user's guide in formulating and encoding queries addressed to the model operational CIDS. The report summarizes the overall process of retrieval from a user's standpoint, and discusses the efficient use of each of the various components of the search system. Numerous illustrations of queries encoded on the standard coding form are provided, along with explanations of the strategy employed in each search. The command language for inputting and processing queries from the teletype is included as an appendix.( )

14. KEY WORDS

Query encoding	Query coding form
Search efficiency	Abnormalities
Molecular formula keys	Nonstructural data keys
Structural fragment keys	Keydesignations
Molecular formula statement	Structurespecifications
Atom-by-atom search	KEYS logical expression